An intelligent decision support system for estimating supply lead times towards improved safety stock dimensioning

Júlio Dinis Lopes de Barros

UMinho | 2023

Universidade do Minho
Escola de Engenharia

Júlio Dinis Lopes de Barros

An intelligent decision support system for estimating supply lead times towards improved safety stock dimensioning

January 2023

FCT Fundação para a Ciência e a Tecnologia

REPÚBLICA PORTUGUESA

CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

NORTE2020
PROGRAMA OPERACIONAL REGIONAL DO NORTE

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Social Europeu

Universidade do Minho
School of Engineering

Júlio Dinis Lopes de Barros

An intelligent decision support system for estimating supply lead times towards improved safety stock dimensioning

Doctoral Thesis
Doctoral Program in Advanced Engineering
Systems for Industry (AESI)

Work developed under the supervision of:
Professor Paulo Cortez
Professor Maria do Sameiro Carvalho

January, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

________________________, ______________________________________
(Place) (Date)

____
(Júlio Dinis Lopes de Barros)

COPYRIGHT AND TERMS OF USE OF THIS WORK BY A THIRD PARTY

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.10.17) (Lourenço, 2021).

Acknowledgements

At the end of this important but also a long journey of my life, I would like to thank all the people who in some way contributed to the completion of this doctoral thesis, to whom I convey my sincere thanks.

First of all, a special thanks to my supervisor Paulo Alexandre Ribeiro Cortez and co-supervisor Maria do Sameiro Carvalho for your enthusiasm, motivation, dedication, guidance, patience, and all the support given during these years. It was a great pleasure working with you as a Doctor of Philosophy (PhD) student.

To Bosch Automotive Electronics, Portugal (AE/P), for providing the conditions for conducting this thesis. A very special thanks to João N. C. Gonçalves for his unceasing availability, help and support. I would also like to thank all the professors and colleagues of the PhD Program in Advanced Engineering Systems for Industry (AESI).

A special thanks to Professor Ricardo Machado for believing in me and opening the world of scientific research to me. To professor Maribel Y. Santos for the motivation, inspiration and friendship.

On a more personal level, a deep and sincere thanks to my family, especially my parents Venâncio Barros and Maria Teresa Barros for their trust and unconditional support. To my brothers/sisters Lucindo Barros, Celestino Barros, Felismino Barros, Auriza Barros, Edmilson Barros and Sandra Barros for the love. To Elisabete Antunes for always being by my side during these last 11 years and for being my safe port. Thank you for the love, patience and understanding during this journey.

I must also thank all my friends and colleagues who accompanied me throughout this great journey. However, I must give special thanks to my long-time friends, Mário Évora, Elisa Furtado and Maria de Fátima Tavares for always being with me. To my lovely friends Nuno Silva, Sofia Moreira, Carina Andrade, Carlos Costa, Ana M. Trigo, Diogo Ribeiro, Rui Oliveira, Pedro Oliveira, Péricles Hungria, Nilton Gomes and Paula Lopes for their friendship, sharing, motivation and trust. My sincere thanks for everything.

This work has been supported by **FCT** – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020 and the doctoral scholarship grant: PD/BDE/142895/2018. Moreover, I would like to thank to Lourenço, 2021 for providing the *novathesis* LaTeX template for free.

"If I have seen further it is by standing on the shoulders of Giants."

— Isaac Newton

"The greatest glory in living lies not in never falling, but in rising every time we fall."

— Nelson Mandela

"In the great battles of life, the first step to victory is the desire to win."

— Mohandas Karamchand Gandhi

Resumo

Um sistema inteligente de apoio à decisão para estimar o lead time para melhorar a estimativa do stock de segurança

O stock de segurança desempenha um papel crucial na manutenção do equilíbrio entre o excesso de inventário e a perda de vendas, o que leva a um melhor desempenho da cadeia de abastecimento. É adotado pelas organizações para cobrir tanto a incerteza da procura como a do Lead-time (LT), a fim de atingir o nível de serviço prometido aos clientes e evitar a ruptura de stock. A incerteza do LT de fornecimento representa um parâmetro central que afeta o dimensionamento do stock de segurança e o desempenho da cadeia de abastecimento. Abordagens baseadas em dados, tais como Big Data Analytics (BDA), têm sido cada vez mais exploradas na gestão da cadeia de abastecimento para a melhoria do processo de tomada de decisão logísticos.

Esta tese propõe um sistema inteligente de apoio à decisão para estimar o LT para um melhor dimensionamento do stock de segurança utilizando uma arquitetura escalável de Big Data (BD). Foca-se na melhoria da estimativa do LT para promover um melhor dimensionamento do stock de segurança.

O estado da arte do dimensionamento do stock de segurança sob incertezas e riscos retrata as dificuldades em determinar as varianções do LT do fabricante numa cadeia de abastecimento com múltiplos produtos. O preenchimento desta lacuna identificada é de grande relevância e motivou-nos a realizar este trabalho. Assim, este projeto propõe duas principais abordagens: 1. uma nova abordagem baseada em aprendizagem automática para prever o risco de atraso de fornecimento, o que representa o principal fator com impacto no LT e no desempenho global da gestão de inventário. 2. uma abordagem supervisionada multivariada para estimar o *lead time* de forma a melhorar as estimativas de stock de segurança, combinando ténicas de aprendizagem automática e de BD. Estas abordagens foram testadas no contexto da Bosch AE/P e reveleram-se muito úteis no suporte à tomada de decisão, na melhoria do desempenho do sistema de gestão de inventários e na gestão proativa e dinâmica dos atrasos de fornecimento.

Palavras-chave: Incerteza do *Lead time*, Atraso do fornecedor, Stock de segurança, Aprendizagem automática, *Big Data*.

Abstract

An intelligent decision support system for estimating supply lead times towards improved safety stock dimensioning

Safety stock plays a crucial role in maintaining the balance between excess inventory and lost sales, which leads to better supply chain performance. It is adopted by organizations to cover both demand and LT uncertainty in order to achieve the promised service level to the customers and prevent stock-outs. Uncertainty in supply LT represents a core parameter that affects the dimensioning of safety stock and supply chain performance. Data-driven approaches such as BDA have been increasingly explored in supply chain management for the enhancement of the logistics decision-making process.

This thesis proposes an intelligent decision support system to estimate supply LT for improved safety stock dimensioning by using a scalable BD technology architecture. It focuses on improving LT uncertainty estimation to promote better safety stock dimensioning.

The state-of-the-art of safety stock dimensioning under uncertainties and risks portrays the difficulties of determining upstream manufacturer's variations of supply LT in the supply chain with multiple products. Fulfilling this identified gap is of major relevance and motivated us to conduct this work. Thus, this work proposes two main approaches: 1. a novel machine learning-based approach for predicting the risk of supply delay, which represents the main factor impacting supply LT and overall inventory management performance. 2. a multivariate supervised approach to estimate supply LT towards the improvement of safety stock estimations, combining machine learning and BD techniques. These approaches were tested in the context of the Bosch AE/P and proved very useful in supporting decision-making, improving the performance of the inventory management system and supply delays management proactively and dynamically.

Keywords: Lead time uncertainty, Supplier delay, Safety stock, Machine learning, Big Data.

Contents

List of Figures xiii

List of Tables xvi

Acronyms xviii

Symbols xxiv

1 Introduction 1

- 1.1 Motivation 2
- 1.2 Research Objectives 3
- 1.3 Research Methodology 3
  - 1.3.1 Design Science Research Methodology for Information System 4
  - 1.3.2 Cross-Industry Standard Process for Data Mining 7
- 1.4 Contributions 9
- 1.5 Document Structure 10

2 Background 13

- 2.1 Industry 4.0 15
  - 2.1.1 Overview of the industrial revolution 15
  - 2.1.2 Industrial Internet of Things 15
  - 2.1.3 Industrial Cyber-Physical Systems 17
- 2.2 Decision Support Systems and Business Analytics 18
  - 2.2.1 Decision Support Systems (DSS) 18
  - 2.2.2 Machine Learning and Predictive Analytics 20

2.3 Business Analytics and Industry 4.0 . . . 22
2.4 Supply Chain and Logistics . . . 23
2.4.1 Procurement . . . 26
2.4.2 Production . . . 27
2.4.3 Distribution . . . 28
2.4.4 Customer Interface . . . 29
2.5 Logistics Processes at Bosch Automotive Electronics . . . 30
2.5.1 Demand planning . . . 33
2.5.2 Production planning . . . 33
2.5.3 Procurement planning . . . 34

3 State of Art 35

3.1 Introduction . . . 37
3.2 Theoretical background . . . 39
3.2.1 Procurement: sourcing and purchasing . . . 39
3.2.1.1 Sourcing . . . 39
3.2.1.2 Purchasing . . . 40
3.2.2 Supply chain risk and uncertainty . . . 40
3.2.3 Uncertainty factors and risks in the procurement process . . . 41
3.2.4 Dimensioning of safety stock (Traditional strategies) . . . 43
3.3 Review methodology . . . 46
3.3.1 Searching phase . . . 46
3.3.2 Selecting phase . . . 47
3.3.3 Analysing phase . . . 47
3.4 Descriptive and co-occurrence analysis . . . 48
3.4.1 Descriptive analysis . . . 48
3.4.1.1 Year of publication . . . 49
3.4.1.2 Venue of publication . . . 49
3.4.1.3 Research problem . . . 49
3.4.1.4 Authors influence and affiliations . . . 50
3.4.1.5 Approach followed . . . 52
3.4.1.6 Research method . . . 52
3.4.2 Co-occurrence analysis . . . 53
3.5 Literature analysis (Scientific contributions) . . . 54
3.5.1 Safety stock dimensioning . . . 54
3.5.1.1 Considering demand uncertainty . . . 54
3.5.1.2 Considering lead time uncertainty . . . 57
3.5.1.3 Considering yield uncertainty . . . 59

3.5.1.4 Considering multiple uncertainties and risks 60
3.5.2 Safety stock management 62
3.5.2.1 Considering demand uncertainty 62
3.5.2.2 Considering lead time and yield uncertainty 65
3.5.2.3 Considering multiple uncertainties and risks 65
3.5.3 Safety stock allocation, positioning or placement 67
3.5.3.1 Considering demand and lead time uncertainty 68
3.5.3.2 Considering multiple uncertainties and risks 72
3.6 Literature gaps and research opportunities 74
3.7 Summary and directions for future research 77

4 Supervised learning of supply delay risk 80
4.1 Advancing Logistics 4.0 with the implementation of a Big Data Warehouse: A Demonstration Case at the Automotive Industry 82
4.1.1 Introduction 82
4.1.2 Related work 84
4.1.3 Propose Architecture for a Logistics 4.0 Big Data Warehouse 86
4.1.3.1 Logical Architecture 86
4.1.3.2 Technological Architecture 88
4.1.4 Demonstration Case 90
4.1.5 Results and Discussion 94
4.1.5.1 Efficacy and Efficiency 94
4.1.5.2 Challenges 96
4.1.5.3 Opportunities 98
4.1.6 Summary and Future Work 100
4.2 A Machine Learning-based framework for predicting supply delay risk using Big Data technology 102
4.2.1 Introduction 103
4.2.2 Aims and contributions 104
4.2.3 Related work 105
4.2.4 Industrial Case Study 107
4.2.5 Material and methods 108
4.2.5.1 Big Data technological architecture 108
4.2.5.2 Machine Learning pipeline 111
4.2.5.3 Data Preparation 111
4.2.5.4 Classification algorithms 116
4.2.5.5 Hyperparameter Tuning 118
4.2.5.6 Evaluation 119

4.2.6 Experiments and Results . . . 123
4.2.6.1 Experimental and modeling setup . . . 123
4.2.6.2 Overall predictive performance . . . 124
4.2.6.3 Misclassification impact on inventory management performance 128
4.2.7 Summary . . . 132

5 Supervised learning for estimating lead-time uncertainty 134
5.1 Introduction . . . 136
5.1.1 Motivation . . . 136
5.1.2 Research objectives and contributions . . . 138
5.2 Related work . . . 139
5.3 Problem formulation . . . 142
5.4 Material and methods . . . 143
5.4.1 Industrial case study . . . 143
5.4.2 Big Data technological architecture . . . 144
5.4.3 Machine Learning pipeline . . . 145
5.4.4 Data Preparation . . . 146
5.4.4.1 Logistics data . . . 146
5.4.4.2 Feature engineering . . . 148
5.4.5 Regression algorithms . . . 150
5.4.5.1 Decision Tree (DT) regression . . . 151
5.4.5.2 Random Forest (RF) regression . . . 151
5.4.5.3 Linear Regression (LR) . . . 152
5.4.5.4 Generalized Linear Model (GLM) . . . 152
5.4.5.5 Gradient-Boosted Tree (GBT) regression . . . 152
5.4.6 Feature preparation and hyperparameters tuning . . . 153
5.4.6.1 Feature scaling . . . 153
5.4.6.2 Feature selection . . . 153
5.4.6.3 Hyperparameters tuning . . . 153
5.4.7 Evaluation . . . 154
5.4.7.1 Measuring model performance . . . 155
5.4.7.2 Measuring model prediction bias . . . 156
5.4.7.3 Safety stock estimation . . . 156
5.5 Experiments and Results . . . 157
5.5.1 Experimental and modeling setup . . . 157
5.5.2 Preliminary experiments . . . 158
5.5.3 Overall predictive performance . . . 159
5.5.4 Safety stock performance . . . 160

5.5.5 Practical & managerial implications 164
5.6 Summary 165

6 Conclusions 167
6.1 Summary 168
6.2 Discussion 169
6.3 Future Work 172

Bibliography 176

List of Figures

1 Design Science Research Methodology (DSRM) Process Model adapted from Peffers et al. (2007). 5

2 Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology adapted from Chapman et al. (2000). 7

3 Generic tasks (**bold**) and outputs of the CRISP-DM methodology adapted from Chapman et al. (2000). 9

4 Overview of industrial revolutions adapted from (Drath & Horch, 2014) 16

5 Industrial Internet of Things (I-IoT) Architecture adapted from (H. Xu et al., 2018) 17

6 Industrial Cyber-Physical Systems (I-CPS) Architecture adapted from (H. Xu et al., 2018) 18

7 Overview of the evolution of Decision Support Systems (DSS) adapted from (Arnott & Pervan, 2016) 20

8 Gartner's Value Chain Model of Analytics adapted from (Koch et al., 2015) 21

9 Machine Learning (ML) categories and respective methods adapted from (Turban et al., 2011) 22

10 Supply Chain adapted from (Council of Supply Chain Management Professional - CSCMP, 2013b) 25

11 Logistics Management adapted from (Council of Supply Chain Management Professional - CSCMP, 2013b) 25

12 Steps of procurement process adapted from (Carvalho et al., 2017) 27

13 Operations Management process adapted from (Rushton et al., 2014) 28

14 Distribution strategy process adapted from (Ghiani et al., 2004) 29

15 Process chain map of Business sector Mobility Solutions (BBM). 31

16 Logistics management value chain. 31

17 Supply Chain Operation Reference (SCOR) model adopted at Bosch. 32

18 Overview of procurement, production and demand planning processes. 32
19 International Production Network (IPN) planning process. 33

20 Steps of Procurement process adapted from (Carvalho et al., 2017) 41
21 Adopted review methodology 48
22 Distribution of scientific publication over the years 49
23 Distribution of publication and their percentage per journal 50
24 Distribution of publications for each safety stock research problem 50
25 Distribution of publications for each safety stock research problem over the years 51
26 Distribution of publication for each safety stock problem under different uncertainty factors and risks 51
27 Geographical locations of authors affiliations (using gpsvisualizer.com) 52
28 Distribution of adopted approaches 53
29 Distribution of research methods 53
30 Co-occurrence map (Level 1) 54
31 Co-occurrence map (Level 3) 54
32 Supply chain structure: a) Serial network, b) Divergent network, c) Convergent network, d) General acyclic network adapted from (H. Li & Jiang, 2012) 68
33 Distribution of Service level measure adopted 76
34 Literature Map 77

35 Logical architecture. 87
36 Technological architecture. 89
37 Development process. 91
38 Big Data Warehouse (BDW) Data Model. 93
39 Implementation Process. 94
40 Country table in Hive. 95
41 Dashboard with historical and predicted data related with deliveries. 96
42 CRISP-DM methodology adapted from Chapman et al. (2000). 106
43 Big Data technological architecture. 110
44 Machine learning pipeline. 112
45 Distribution of output target. 113
46 Rolling window scheme adapted from (Oliveira et al., 2017) 119
47 Confusion matrix for a binary classification task. 120
48 Cost matrix for the binary classification task. 122
49 Evolution of Area Under the ROC Curve (AUC) score over the Rolling Window (RW) iterations. 127
50 Random Forest (RF) Receiver Operating Characteristic (ROC) curve at iteration $U = 19$ and selected decision threshold $D = 0.340$. 127

51 Confusion matrix of the RF model at RW iteration *U = 20* when using *D* = 0.340. . . . 127
52 Feature importance for the RF model. 128
53 PowerBI Dashboard: Supply delay risk prediction. 131

54 Illustration of a classical three-stage supply chain structure. 143
55 Geographic distribution of the Bosch AE/P suppliers. 144
56 Big Data technological architecture proposed in N. Silva et al. (2021) 145
57 Machine learning pipeline. 146
58 (left) LT and (right) $log(LT + 1)$ distributions with Kernel Density Estimates (KDE) . . . 148
59 Hyperparameter tuning using Bayesian Optimization. 154
60 Rolling window scheme adapted from (Oliveira et al., 2017) 155
61 MAE and elapsed time (in seconds) as a function of the number of features included in the regression models. 159
62 Evolution of Mean Absolute Error (MAE) metric (*y*-axis) over the RW iterations for the different ML regression models (*x*-axis). 161
63 Coefficient of determination and Regression Error Characteristic (REC) curve at RW iteration *U=5* of the RF model. 161
64 Measured vs Predicted Lead-time at RW iteration *U=5* of RF model. 162
65 Measured vs Predicted Lead-time at RW iteration *U=20* of RF model. 162
66 Feature importance for the RF model. 162
67 Dashboard: Lead-time prediction and safety stock estimation. 163
68 Safety stocks evolution as a function of target service levels. 163

69 Intelligent Decision Support System for safety stock dimensioning. 175

List of Tables

1 Risk definitions 41

2 Supply Chain Risks (SCR) definitions 41

3 Traditional methods for safety stock dimensioning (Schmidt et al., 2012; Yamazaki et al., 2016) 45

4 Query for searching of bibliographic references (Literature analysis) 47

5 Key contributing authors (first author) 51

6 Top contributing affiliations 52

7 Top adopted techniques 53

8 Chronological scientific contributions on safety stock dimensioning under demand uncertainty 58

9 Chronological scientific contributions on safety stock dimensioning under lead time uncertainty 59

10 Chronological scientific contributions on safety stock dimensioning under yield uncertainty 60

11 Chronological scientific contributions on safety stock dimensioning under multiple uncertainties and risks 61

12 Chronological scientific contributions on safety stock management under demand uncertainty 63

13 Chronological scientific contributions on safety stock management under lead time and yield uncertainties 65

14 Chronological scientific contributions on safety stock management under multiple uncertainties and risks 66

15 Chronological scientific contributions on safety stock allocation, positioning and placement under demand and lead time uncertainties 70

16 Chronological scientific contributions on safety stock allocation, positioning and placement under demand and lead time uncertainties 71

17 Chronological scientific contributions on safety stock allocation, positioning and placement under multiple uncertainties and risks 73

18 Chronological scientific works for modeling supply delays risk using BDA techniques. . . 108

19 Summary of the logistics data attributes. 113

20 Hyperparameter of the classification algorithms 118

21 Special freight cost functions according to the transportation mode and supplier location. 123

22 Experimental setup. 125

23 Configuration of the *spark-submit* parameters (Spark, 2020). 125

24 Results of feature selection experiments in terms of AUC. 126

25 Comparison of the optimized ML models (best values in **bold**). 126

26 Examples of Misclassifying impact calculation. 129

27 ML models cost (lowest costs in **bold**). 129

28 RF model cost for selected threshold *D* values. 130

29 ML model predictions for selected threshold *D* values 131

30 Chronological scientific works for modeling LT using BDA techniques 141

31 Summary of the collected logistics data attributes. 147

32 Hyperparameter of the regression algorithms 151

33 Comparison of the optimized ML models (best values in **bold**). 159

34 Models prediction bias (best values in **bold**). 161

35 Deviations in safety stocks (in units) with respect to the company estimation. 164

36 Deviations in safety stock costs with respect to the company estimation. 164

Acronyms

R^2	Coefficient of Determination *(pp. 8, 160)*
sME	scaled Mean Errors *(p. 156)*
$sMSE$	scaled Mean Squared Error *(p. 156)*
sSE	scaled Squared Errors *(p. 156)*

ABC	Artificial Bee Colony *(p. 64)*
AE/P	Automotive Electronics, Portugal *(pp. iv, vi, vii, 3, 5, 6, 12, 13, 80, 134, 138, 142–144, 146, 171, 172)*
AHR	Adaptive Heuristic Critic *(p. 21)*
AI	Artificial Intelligence *(pp. 20, 82, 103, 104, 173)*
AM	Additive Manufacturing *(p. 82)*
ANN	Artificial Neural Network *(pp. 140, 173)*
AOREC	Area Over the REC curve *(p. 156)*
API	Application Programming Interface *(pp. 89, 109, 157)*
ATO	Assemble-to-Order *(p. 55)*
ATP	Available-To-Promise *(p. 60)*
ATPE	Adaptive Tree of Parzen Estimators *(p. 119)*
AUC	Area Under the ROC Curve *(pp. xiv, 8, 104, 106, 116, 119–121, 124, 126, 127, 132)*
AUREC	Area under the Regression Error Characteristic *(pp. 8, 156, 160)*
AutoML	Automated Machine Learning *(p. 173)*

BA	Business Analytics *(pp. 13, 19, 22, 23, 74–76, 168)*
BBM	Business sector Mobility Solutions *(pp. xiii, 30, 31, 33)*

BD	Big Data *(pp. vi, vii, 8, 82, 87, 88, 107, 136, 138, 142, 145, 170, 172)*
BDA	Big Data Analytics *(pp. vi, vii, xvii, 11, 12, 74–76, 86, 90, 102–105, 107, 108, 138, 139, 141, 144, 169)*
BDW	Big Data Warehouse *(pp. xiv, 11, 82, 84–88, 90–94, 98, 100, 101, 104, 142, 146, 165, 171)*
BFGS	Broyden–Fletcher–Goldfarb–Shanno *(p. 118)*
BI	Business Intelligence *(p. 19)*
BI-A	Business Intelligence and Analytics *(p. 19)*
BPS	Bosch Production System *(p. 32)*
CART	Classification And Regression Tree *(p. 151)*
CPS	Cyber-Physical Systems *(pp. 15–17, 82, 103)*
CRISP-DM	Cross-Industry Standard Process for Data Mining *(pp. xiii, 3, 4, 7, 9, 105)*
CSCMP	Council of Supply Chain Management Professional *(p. 23)*
CSL	Cycle Service Level *(pp. 45, 46, 75)*
CST	Constant Service Time *(p. 69)*
DDMRP	Demand-Driven Materials Requirement Planning *(p. 57)*
DGI	Decrease Gini Impurity *(p. 151)*
DL	Deep Learning *(p. 173)*
DM	Data Mining *(pp. 7, 8)*
DPSS	Dynamic Planned Safety Stock *(p. 60)*
DSRM	Design Science Research Methodology *(pp. xiii, 4, 5, 84)*
DSRM-IS	Design Science Research Methodology for Information Systems *(pp. 3, 4, 7)*
DSS	Decision Support Systems *(pp. xiii, 13, 18–20, 138)*
DT	Decision Tree *(pp. 8, 116, 124, 126, 129–132, 140, 150, 151, 157–160, 165)*
DW	Data Warehouse *(pp. 83–86)*
EDA	Exploratory Data Analysis *(pp. 7, 112, 147)*
EIS	Executive Information Systems *(p. 19)*
ELSP	Economic Lot Scheduling Problem *(p. 55)*
ELT	Extract, Load and Transform *(p. 84)*
EOQ	Economic Order Quantity *(pp. 59, 62, 63, 67)*
ERP	Enterprise Resource Planning *(pp. 22, 78, 90, 97, 111, 142)*
ETA	Estimated Time of Arrival *(pp. 106, 107)*
ETL	Extract, Transform and Load *(p. 84)*
FN	False Negatives *(pp. 120, 128, 130)*

FNR	False Negative Rate *(p. 8)*
FOP	Forecast Order Policy *(p. 64)*
FP	False Positives *(pp. 120, 128, 130)*
FP	Fokker Planck *(p. 64)*
FPR	False Positive Rate *(p. 121)*
FR	Fill Rate *(pp. 45, 46, 75)*
GA	Genetic Algorithm *(p. 62)*
GARCH(1,1)	Generalized Autoregressive Conditional Heteroscedastic *(p. 57)*
GB	gigabytes *(p. 157)*
GBT	Gradient-Boosted Tree *(pp. 8, 116, 117, 124, 126, 129, 130, 132, 150, 152, 157–160, 165)*
GLM	Generalized Linear Regression *(pp. 8, 150, 152, 157–160, 165)*
GPS	Global Positioning Systems *(p. 22)*
HDFS	Hadoop Distributed File System *(pp. 86, 88, 89, 91, 92, 94, 108–111, 144, 146)*
I-CPS	Industrial Cyber-Physical Systems *(pp. xiii, 13, 15, 18, 22)*
I-IoT	Industrial Internet of Things *(pp. xiii, 13, 15–18, 22)*
IB-ISP	Inventory-Theory-Based Interval Stochastic Programming *(pp. 66, 67)*
IBC	Intermediate Bulk Container *(p. 29)*
IDSS	Intelligent Decision Support System *(pp. 3, 11, 19, 136, 138, 139, 142, 144, 165, 169, 171, 172)*
IFTN	Intermodal Freight Transport Networks *(p. 106)*
IHC	Inventory Holding Costs *(p. 121)*
iid	independently and identically *(p. 67)*
IoT	Internet of Things *(pp. 15, 16, 82, 103)*
IPN	International Production Network *(pp. xiv, 33)*
IS	Information System *(pp. 4, 18)*
IT	Information Technology *(pp. 4, 103, 172)*
JIT	Just-In-Time *(pp. 2, 57, 138, 140, 143)*
JSON	JavaScript Object Notation *(p. 92)*
K-NN	K-Nearest Neighbors *(p. 106)*
L-BFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno *(pp. 117, 118, 158)*
LightGBM	Light Gradient Boosting Machine *(p. 173)*

Lime	Local Interpretable Model-Agnostic Explanations *(p. 173)*
LM	Logistics Management *(p. 23)*
LPM	Logistics Planning Management *(pp. 30, 32)*
LR	Linear Regression *(pp. 8, 150, 152, 158–160)*
LR	Logistic Regression *(pp. 8, 116, 117, 124, 126, 129–132, 140, 157, 160, 165)*
LSVC	Linear Support Vector Machine *(pp. 116, 117, 124, 126, 132)*
LT	Lead-time *(pp. vi, vii, xvii, 2, 3, 11, 12, 137–142, 144, 147, 152, 156, 157, 160, 162, 165, 166, 168–172, 174)*
M2M	Machine-to-Machine *(p. 17)*
MAE	Mean Absolute Error *(pp. xv, 8, 154–156, 158–161)*
MILP	Mixed-Integer Linear Programming *(pp. 61, 63)*
MINLP	Mixed-Integer Non-Linear Programming *(p. 63)*
MIP	Mixed-Integer Programming *(p. 64)*
ML	Machine Learning *(pp. xiii, xv, xvii, 3, 6, 8, 11, 12, 20–23, 94, 102–109, 111, 112, 114, 117, 118, 123, 124, 126, 129, 131–133, 136, 138–140, 142, 144–146, 153, 157, 159, 161, 165, 169–173)*
MLP	Multilayer Perceptron *(pp. 8, 116–118, 124, 126, 129, 132)*
MODE/D	Decomposition-based Multi-Objective Differential Evolution algorithm *(pp. 61, 67)*
MOEMO	Multi-Objective Electromagnetism-like Optimization *(p. 63)*
MOPSO	Multi-Objective Particle Swarm Optimization *(pp. 56, 63)*
MPP	Massively Parallel Processing *(p. 89)*
MPP	Massively Parallel Processing *(p. 110)*
MPS	Master Production Schedule *(pp. 28, 60)*
MRP	Material Requirements Planning *(pp. 12, 28, 34, 42, 54, 55, 60, 62, 65, 75, 78, 90, 138, 142, 143, 160)*
MSE	Mean Squared Error *(p. 8)*
MTO	Make-to-Order *(pp. 55, 59, 140)*
NBDRA	NIST Big Data Reference Architecture *(p. 85)*
NoSQL	Not Only SQL *(pp. 85, 88)*
NPV	Negative Predictive Value *(p. 120)*
NPV	Net Present Value *(p. 55)*
NSF	National Science Foundation *(p. 17)*
OEM	Original Equipment Manufacturer *(pp. 30, 32, 33)*
OLAP	Online Analytical Processing *(p. 19)*

OM	Operations Management *(p. 27)*
OR	Operation Research *(p. 140)*
OWL-QN	Orthant-Wise Limited-memory Quasi-Newton *(p. 117)*
PhD	Doctor of Philosophy *(pp. iv, 3, 6, 9, 10, 172)*
PLC	Product Life Cycle *(pp. 64, 76, 78, 174)*
PO	Purchase Order *(p. 26)*
POT	Power-of-Two *(p. 64)*
PPV	Positive Predictive Value *(p. 120)*
RCCP	Rough Cut Capacity Planning *(p. 28)*
REC	Regression Error Characteristic *(pp. xv, 156, 160, 161)*
RF	Random Forest *(pp. xiv, xv, xvii, 8, 102, 116, 117, 124, 126–132, 136, 140, 150, 151, 157–161, 165, 173)*
RFID	Radio Frequency-based Identification *(pp. 17, 22)*
RFQ	Request for Quotation *(p. 26)*
RMSE	Root Mean Squared Error *(pp. 8, 140, 155, 156, 158, 159)*
RO	Research Objectives *(pp. 3, 9, 10)*
ROC	Receiver Operating Characteristic *(pp. xiv, 8, 104, 120, 121, 124, 126)*
RS	Random Search *(p. 119)*
RW	Rolling Window *(pp. xiv, xv, 11, 105, 119, 120, 123, 124, 126–129, 132, 139, 154, 157–161, 165)*
SAP	System Analysis Program Development *(p. 61)*
SARSA	State-Action-Reward State-Action *(p. 21)*
SC	Supply Chain *(pp. 2, 10–13, 30, 102, 103, 136, 137, 140, 166, 168, 169)*
SCLSP	Stochastic Capacitated lot-sizing Problem *(p. 57)*
SCM	Supply Chain Management *(pp. 2, 23, 24, 76, 168)*
SCOR	Supply Chain Operation Reference *(pp. xiii, 30, 32)*
SCR	Supply Chain Risks *(pp. xvi, 41)*
SCRM	Supply Chain Risk Management *(pp. 103, 104)*
SDN	Software-Defined Networking *(p. 17)*
SDT	Stock Diffusion Theory *(p. 64)*
SHAP	SHapley Additive exPlanations *(p. 173)*
SJR	SCImago Journal Rank *(p. 47)*
SKU	Stock Keeping Units *(p. 42)*
SL	Service Level *(pp. 12, 45)*

SLR	Systematic Literature Review *(pp. 3, 5, 35, 37, 38, 47, 49, 57, 65, 67, 73, 74, 169)*
SOQ	Sustainable Order Quantity *(p. 59)*
SQL	Structured Query Language *(pp. 88, 89, 92, 109–111)*
SS	Safety Stock *(pp. 2, 137, 138, 142, 163, 166, 168, 172)*
SVM	Support Vector Machine *(pp. 117, 140, 173)*
TB	terabytes *(p. 157)*
TN	True Negatives *(p. 120)*
TNR	True Negative Rate *(p. 120)*
TP	True Positives *(p. 120)*
TPE	Tree of Parzen Estimators *(pp. 119, 154)*
TPR	True Positive Rate *(pp. 8, 120)*
TPR	Time Replenishment *(pp. 45, 121)*
TSS	Time to reach the Safety Stock level *(p. 56)*
WoS	Thomson Reuters' Web of Science *(pp. 46, 47, 77)*
WPAN	Wireless Personal Area Networks *(p. 17)*
WSAN	Wireless Sensor and Actuator Network *(p. 17)*
XAI	Explainable Artificial Intelligence *(p. 173)*
XGBoost	eXtreme Gradient-Boosting Machine *(p. 173)*

Symbols

α	Cycle Service Level *(p. 45)*
β	Fill Rate *(p. 45)*
μ_D	Mean of demand *(p. 157)*
μ_{LT}	Mean of lead time *(p. 157)*
$\Phi(\cdot)$	Standard Gaussian cumulative distribution function *(p. 157)*
σ_1	One-step-ahead standard deviation of the forecast error *(p. 45)*
σ_D	Standard deviation on-demand *(p. 157)*
σ_F	Standard deviation of the forecast error for the demand during TRP *(p. 45)*
σ_L	Standard deviation of the forecast error for a certain lead time L *(p. 45)*
σ_{LT}	Standard deviation of lead time *(p. 157)*

1 Introduction

Summary: This first chapter initiates with the motivation that led to addressing the research problem for this doctoral thesis. Afterwards, the research objectives are outlined, followed by the research methodology which conducts the research process. Moreover, all scientific contributions produced are listed. Lastly, this chapter closes with a description of the structure of this document.

Chapter Table of Contents:

1.1 Motivation . . . 2
1.2 Research Objectives . . . 3
1.3 Research Methodology . . . 3
1.3.1 Design Science Research Methodology for Information System . . . 4
1.3.2 Cross-Industry Standard Process for Data Mining . . . 7
1.4 Contributions . . . 9
1.5 Document Structure . . . 10

1.1 Motivation

The Supply Chain (SC) is a complex and unique network that integrates different business processes involved in fulfilling customer needs, that includes production, warehouses, retailers, transportation, suppliers and even customers (Chopra & Meindl, 2016). All these elements are involved in the entire product life cycle, from procurement to manufacturing, distribution and customer service (Balfaqih et al., 2016). The importance of the Supply Chain Management (SCM) in business strategy, in attracting and retaining customers and markets, in the effectiveness of operation management and the profitability of companies has become a valuable way to ensure the competitive advantage and improve the organisational performance (Balfaqih et al., 2016; Carvalho et al., 2017; Trkman et al., 2010). Today's supply chains are characterised by being a very complex network and more exposed to uncertainties and risks, caused mainly by the influence of globalisation and global market competitiveness. In particular, the recent and current pandemic of Coronavirus (SARS-CoV-2) has made the SC even more exposed to these uncertainties and risks, causing severe supply and demand problems in SC network, affecting both the customer satisfaction levels and supply chain-related costs negatively.

The automotive industry has incorporated the Just-In-Time (JIT) concept, which is concerned with demand-driven production to reduce overall waste, particularly inventory levels. Thus, adopting this philosophy allows for reducing inventory on-hand, yet, becoming even more vulnerable to uncertain occurrences (Vieira et al., 2019). Therefore, managing these uncertainties and risks becomes a fundamental challenge for such SCs. MRP buffering techniques such as Safety Stock (SS) are adopted to cover these uncertainties to achieve the promised customer service level and avoid stock-outs. SS is crucial to maintain the balance between excess inventory and lost sales.

Customer demand and supply LT uncertainties are two core parameters to determine safety stock. Thus, it is common sense that these uncertainties can cause stockouts or supply chain disruptions. However, researchers have paid more attention to demand uncertainty rather than to supply LT uncertainty (Dolgui & Prodhon, 2007; Heydari et al., 2009). Conceptually, supply LT uncertainty remains a core parameter that varies and affects not only the inventory management but also SC performance (Chaharsooghi & Heydari, 2010; B. Dey et al., 2021; Heydari et al., 2009; Z. Li et al., 2019).

Most of the traditional inventory models have assumed supply LT as deterministic or normally distributed stationary, which is not realistic in the real-world supply chain due to the random events that cause delays. Such delays may require incurring in special/premium freights to avoid stockouts and consequently an extra cost for organizations. This statistical assumption of the normal distribution can lead to a higher service level than desired, resulting in an overestimation of safety stock and consequently higher inventory costs (Ruiz-Torres & Mahmoodi, 2010). Only few literature contributions considered the existence of LT uncertainty issues (see, e.g., Abdel-Malek et al., 2005; Chopra et al., 2004; Digiesi et al., 2013; Disney et al., 2016; Kanet et al., 2010; M. Louly & Dolgui, 2009; Ruiz-Torres & Mahmoodi, 2010; Saad, Perez, & Alvarado, 2017; Talluri et al., 2004).

Our research is motivated by the difficulties of determining upstream (manufacturer's) variations of supply lead-time in supply chains with multiple products. ML techniques to estimate supply LT towards improving the dimensioning of safety stocks has been explored. Data-driven approaches allow to consider other variables that impact the dynamics of a supplier LT, contrarely to the standard and conventional statistical methods. Moreover, these variables allow capturing recent changes in supplier response patterns with more significant flexibility.

1.2 Research Objectives

The main purpose of this thesis is to propose, design, implement and validate an approach to improve the safety stock levels estimations, in order to promote the optimization of the trade-off between holding inventory costs and special/premium freight costs while attending to a certain service level. Specifically, the goal is to develop ML-based approaches to estimate the replenishment time of supply orders and, ultimately, to promote a better estimation of safety stocks. In order to accomplish this purpose, this thesis address the following Research Objectives (RO):

- **RO1** - Development of a Systematic Literature Review (SLR) related to safety stock dimensioning strategies under uncertainties and risks in the procurement process. The main goal is to identify the literature gaps and research opportunities regarding this research topic.

- **RO2** - Design, implement and validate a machine learning-based framework to predict supply delay risk. Supply delays have a direct influence on supplier performance and also represent one of the major causes of long supply lead times, leading to failures in stock replenishment. Anticipating potential delays helps to ensure an appropriate logistics performance and control, allowing to generate cost savings and the optimization of the trade-off between holding inventory costs and special freight costs.

- **RO3** - Design, implement and validate an Intelligent Decision Support System (IDSS) for estimating supply lead times towards improved safety stock dimensioning. This IDSS aims to predict supply lead time using a scalable technological Big Data architecture, focusing on enhancing lead time uncertainty estimation to ultimately promote better safety stock estimations.

1.3 Research Methodology

To perform this PhD thesis was adopted two main methodologies: The Design Science Research Methodology for Information Systems (DSRM-IS) and CRISP-DM. The DSRM-IS was adopted because this work is based on the development and evaluation of an artifact to address or solve an identified organization problem, in that case, a logistics problem faced by Bosch AE/P regarding safety level estimations. On the other hand, the CRISP-DM provides useful guidelines for conducting a real-world data mining project.

The CRISP-DM will be applied in the third step (Design and development) of the DSRM-IS methodology, i.e., it will conduct the design and development step of DSRM-IS. However, there are several overlaps or complementarities in several steps of these two methodologies, namely in steps 1 and 2 of DSRM-IS with step 1 of CRISP-DM, and in step 5 of both methodologies. The following sub-sections will describe these two methodologies.

1.3.1 Design Science Research Methodology for Information System

Two different paradigms can be adopted for research in the Information System (IS) discipline: behavioral science or design science. The behavioral science paradigm consists of developing and verifying theories that explain or predict human or organizational behavior. On the other hand, the design science paradigm consists of achieving the knowledge and understanding of the problem and its solutions by building and application of the Information Technology (IT) artifact (Hevner et al., 2004). "Design science... creates and evaluates IT artifacts intended to solve identified organizational problems"(Hevner et al., 2004; Peffers et al., 2007).

Two design processes and four types of artifacts are produced by using DSRM-IS: constructs, models, methods, and instantiations.

- **Constructs** - consists of a vocabulary and symbols used to define problems and communicate the solutions (Hevner et al., 2004);

- **Models** - are abstractions and representation of the real-world and the constructs are the bases for this representation (Hevner et al., 2004);

- **Methods** - provides guidance to solve the problem, algorithms, and practices. The methods can be formal through mathematical algorithms, or informal method through the textual description of best practice approaches (Hevner et al., 2004);

- **Instantiations** - are the operationalization of the constructs, models, and methods. The instantiation demonstrates feasibility in the design process and designed product (Hevner et al., 2004).

And, the two design processes are: build and evaluate. In the build design process, the artifacts are built to solve a real-world problem; And, in evaluating design process the artifacts are evaluated regarding their utility in solving those problems (Hevner et al., 2004);

The DSRM-IS cover principles, practices, and procedures required to carry out this type of research and attend three objectives: consistent with prior literature, provides a nominal process model for doing Design Science research and provides a mental model for presenting and evaluating Design Science research in IS (Peffers et al., 2007). The DSRM process model consists into six steps/activities: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication, and four different research entry points: problem-centered initiation, objective-centered solution, design and development-centered initiation and client/context initiated

(Peffers et al., 2007), as depicted Figure 1. This process is structured sequentially but is not mandatory to follow the steps in a sequential way. It can start at any step. The problem-centered initiation research starts with step 1 because the idea for research resulted from the problem observation or from future research suggested in a paper from a prior project. The objective-centered solution research starts with step 2, in a way that the research is trigged in the industry due to the problem that they face to or the research is oriented for artifact developing. The design and development-centered initiation start with step 3, and the research result from the existence of an artifact that was developed to solve other problems, in order to identify functionalities and performance requirements for a new artifact. And lastly, the client/context-initiated research starts with step 4 and is based on practical observation of the final solution, this is, consists to identify the impact of the solution (Peffers et al., 2007).

The research entry point of this work is considered to be problem-centered. This doctoral thesis was assigned and conducted at Bosch AE/P company in order to address or solve a real-world problem faced by the organization. As such, the objectives were already pre-defined and aligned with the needs of the organization. However, it was necessary to a priori review the literature in order to align with existing gaps in the literature and research opportunities.

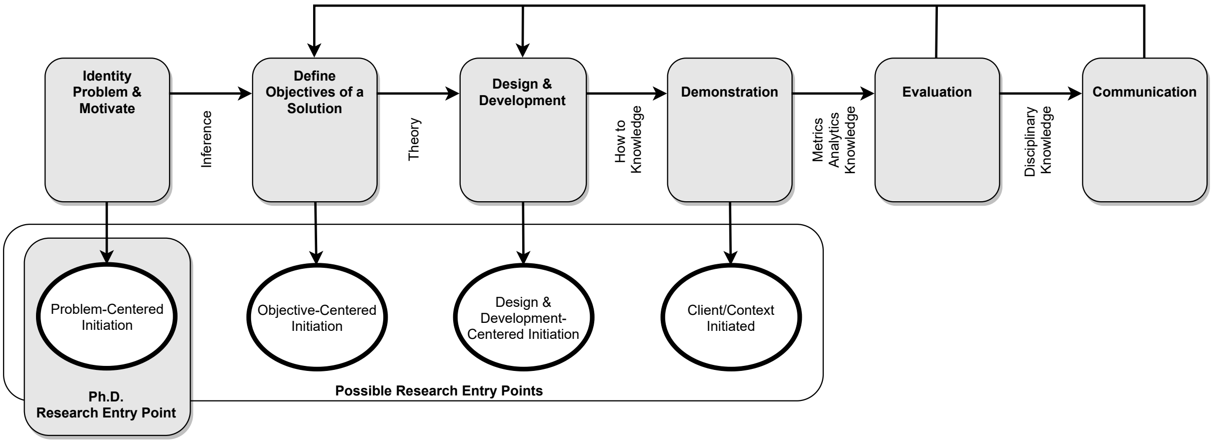

Identity Problem & Motivate
Inference
Define Objectives of a Solution
Theory
Design & Development
How to Knowledge
Demonstration
Metrics Analytics Knowledge
Evaluation
Disciplinary Knowledge
Communication
Problem-Centered Initiation
Objective-Centered Initiation
Design & Development-Centered Initiation
Client/Context Initiated
Ph.D. Research Entry Point
Possible Research Entry Points

Figure 1: DSRM Process Model adapted from Peffers et al. (2007).

A. **Problem identification and motivation**: this step consists to define the research problem and justify the value of a solution (Peffers et al., 2007). At this point, a SLR was employed in order to analyse literature contributions regarding the safety stock problems research under uncertainties and risks in the procurement process, focusing on the dimensioning problem. Moreover, this SLR analysis allowed the identification of literature gaps and research opportunities, providing a road map to guide future research in this topic (see, Chapter 3);

B. **Define the objective for a solution**: consists to define objectives of a solution from the problem definition and the definition of feasible tasks. The objectives can be qualitative or quantitative (Peffers et al., 2007). As aforementioned, the objectives of this thesis are aligned with the needs of the organization. Furthermore, this thesis is also part of a real innovation project. As such, the objectives, purposes and goals are already defined, thus ensuring also the alignment of these objectives, as specified in Section 1.2;

C. **Design and development**: consist of determining the functionalities for the artifact and its architecture, and then create the artifact. The produced artifacts can be constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), instantiations (operationalization of the constructs, models, and methods) or new properties of technical, social, and/or informational resources (Peffers et al., 2007). At this point, two IT artifacts based on ML-based technique were developed in this thesis;

D. **Demonstration**: consists to demonstrate the use of the artifact in order to solve at least one instance of the problem identified in the beginning. This demonstration can be by experimentation, simulation, case study, proof, or other appropriate activity (Peffers et al., 2007). Two industrial case studies are demonstrated in this PhD thesis in order to solve real problems: predicting supply delay risk and lead time variability towards improving safety stock estimations;

E. **Evaluation**: consists to observe and measure the effectiveness of the artifact as the solution for the problem by comparing the objectives of a solution to the actual observed results obtained in the demonstration (using metrics and analysis techniques). The evaluation can be done by comparing the artifact functionality with the solution objectives (defined in the second step), objective quantitative performance measures (for example, budgets, produced items, system availability and response time), the results of the surveys satisfaction, client feedback or simulations (Peffers et al., 2007). The evaluation of the demostration cases aims to observe and measure how well the proposed ML-based approaches works regarding the problems solutions. Besides comparing the proposed artifacts with the solution objectives, several evaluation metrics were also adopted (see Section 1.3.2):

   - **Machine learning metrics** - evaluation of the overall predictive performance of ML models;
   - **Supply chain performance** - evaluation of the proposed ML-based approaches in terms of the inventory-related total costs (inventory holding costs and special freight costs) and advantages of the proposed approach in comparison with the actual used by the case study company.

F. **Communication**: consists to communicate the importance of the problem and the problem properly speaking. Also consists to communicate the artifact and it's utility, novelty, rigor, design, and effectiveness to relevant audiences (Peffers et al., 2007). This activity mainly includes the writing and publishing of this doctoral thesis, and scientific publication in journals, as described in the Section 1.4. In addition, the status/updates of the work carried out were presented in the annual presentations to the management (direction board) of Bosch AE/P.

1.3.2 Cross-Industry Standard Process for Data Mining

CRISP-DM is a methodology that was created in 1996 and in that time the data mining market was new and immature (Chapman et al., 2000). A set of companies including Daimler Chrysler (then Daimler-Benz), SPSS (then ISL), NCR Corporation, formed a consortium to create the standard approach to Data Mining (DM). The result of their work was the CRISP-DM (Chapman et al., 2000; North, 2012).

The CRISP-DM methodology provides an overview of the life cycle of a DM project and consists of six different phases, as depicted in Figure 2.

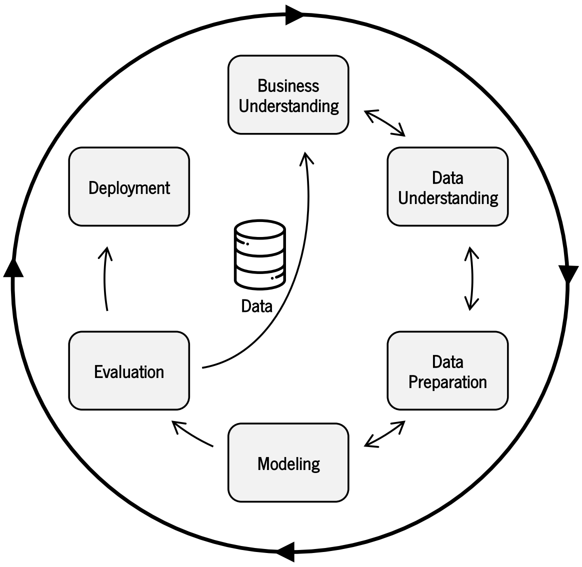


Figure 2: CRISP-DM methodology adapted from Chapman et al. (2000).

A. **Business understanding**: the focus of this phase is to understand the project objectives and requirements from a business perspective, then formulate the data mining problem based on this acquired knowledge and perform a preliminary plan designed to accomplish the objectives. As aforementioned, this phase is related to the first and second step of DSRM-IS (see Section 1.3.1);

B. **Data understanding**: this phase aims to collect data, familiarize with these data (e.g., explore the data) and perceive or verify their quality. Two reports are developed in order to familiarize with dataset, as follows:

- **Data Quality report** - overview of data cardinaly, correlation, missing values and zeros;
- **Data Exploration report** - development of a Exploratory Data Analysis (EDA) in order to explore the dataset and create graphical display of the data.

C. **Data preparation**: this phase is very important to achieve good results. It covers all activities to construct the dataset. The data preparation task includes a table, record, attribute selection, and transformation and cleaning of data that will be used in modeling tools. In this work, the knowledge of the domain expert has become fundamental to performing both feature selection (selection of variables than can impact the problem under consideration from the original dataset) and feature engineering tasks (construction of new features from the original dataset variables, in order to enrich the dataset and therefore used as input from the ML models);

D. **Modeling**: various modeling techniques, such as decision trees, neuronal networks, genetic algorithm, association rules, clustering, and others are selected and applied in this phase, and their parameters are calibrated to optimize the results (Chapman et al., 2000). In DM, the model represents the application of algorithms to search, identify, and display patterns or messages in a given dataset (North, 2012). Focusing in the work developed in this thesis, several ML algorithms were tested to address the binary classification problem (supervised learning) related to the prediction of supply delay risk, such as RF, Logistic Regression (LR), Gradient-Boosted Tree (GBT), Multilayer Perceptron (MLP) and Decision Tree (DT). On the other hand, for the regression problem (supervised learning) regarding to the prediction of the supply lead time were tested RF, Linear Regression (LR), GBT, Generalized Linear Regression (GLM) and DT ML algorithms;

E. **Evaluation**: this phase consists to evaluate the ML model. Before proceeding to the final deployment of the model is important to evaluate it, verifying and analysing that the results meet the business objectives, and review the steps executed to create the model. As aforementioned, we adopted several metrics to evaluate the ML models. The overall performance of classification models (R02) is given by the AUC of ROC analysis, which measures the quality of the probabilistic classifier. A model with an AUC of 50% represents a random classifier, 60% a reasonable classifier, 70% a good classifier, 80% very good classifier, 90% an excellent classifier and 100% a perfect classifier (Ribeiro et al., 2022). Moreover, other classification metrics were also considered, such as True Positive Rate (TPR) and False Negative Rate (FNR). Regarding the regression problem (R03), the overall performance of ML models is given by MAE metric, which provides the difference between the actual value and the model predicted values. Other regression metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2) and Area under the Regression Error Characteristic (AUREC) curve were also considered. As a complement to the regression metrics, model prediction bias was also considered in order to measure how far is the estimated value from the real value. Furthermore, for both classification and regression problems were measured the supply chain performance based on the inventory-related total costs (e.g., inventory holding costs and special freight costs);

F. **Deployment**: the obtained knowledge needs to be organized and presented in order to be used by the customer. Besides that, activities in this phase include generating or producing a report, performing a monitoring and maintenance plan, and reviewing the project (Chapman et al., 2000). The deployment of ML models in BD cluster were configurated by a set of workflow schedules in order to train and re-train the ML models in a pre-defined time frame so that to allow new predictions.

Figure 3 presents an overview of CRISP-DM phases, including the generic tasks and outputs for each phase.

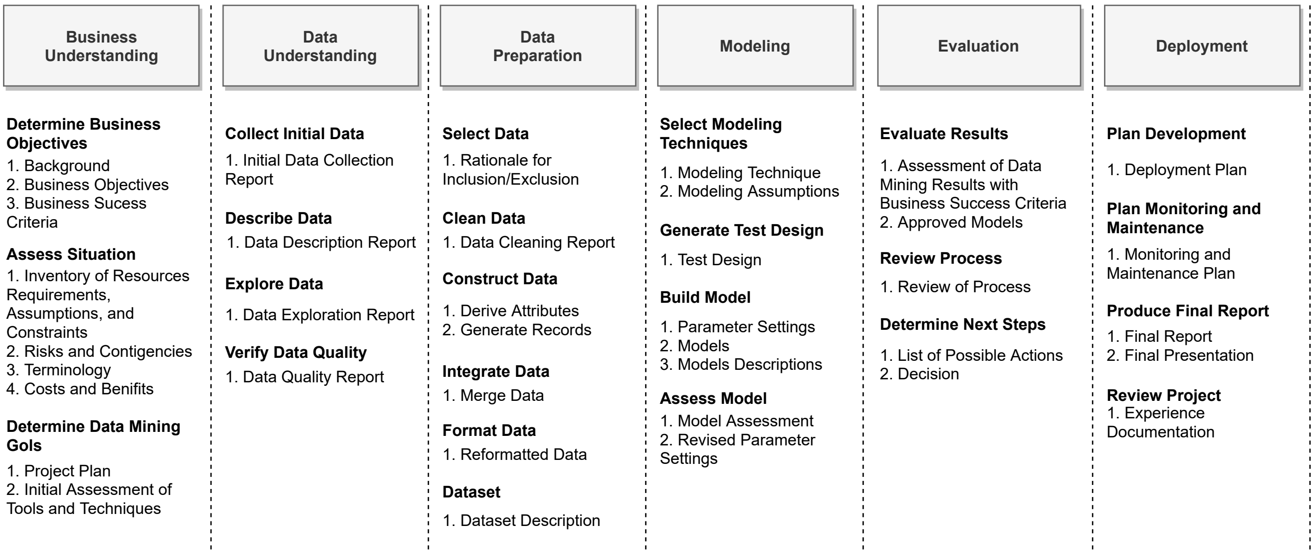


Figure 3: Generic tasks (**bold**) and outputs of the CRISP-DM methodology adapted from Chapman et al. (2000).

1.4 Contributions

This PhD thesis resulted in the publication of 4 scientific publications, which aim to address the objectives outlined, as follows:

- **Title:** A systematic literature review about dimensioning safety stock under uncertainties and risks in the procurement process
  Authors: Júlio Barros, Paulo Cortez and M. Sameiro Carvalho
  Journal: Operations Research Perspectives (ORP)
  DOI: https://doi.org/10.1016/j.orp.2021.100192
  RO: RO1

- **Title:** Advancing Logistics 4.0 with the implementation of a Big Data Warehouse: A Demonstration Case at the Automotive Industry
  Authors: Nuno Silva, Júlio Barros, Maribel Y. Santos, Carlos Costa, Paulo Cortez, M. Sameiro Carvalho and João N.C. Gonçalves
  Journal: Electronics
  DOI: https://doi.org/10.3390/electronics10182221
  RO: RO2

- **Title:** A Machine Learning-based framework for predicting supply delay risk using Big Data technology
  Authors: Júlio Barros, Nuno Silva, João N.C. Gonçalves, Paulo Cortez, M. Sameiro Carvalho, Maribel Y. Santos and Carlos Costa
  Journal: Submitted to a Journal
  RO: RO2

- **Title:** A machine learning strategy for estimating supply lead times towards improved safety stock dimensioning
  Authors: Júlio Barros, João N.C. Gonçalves, Paulo Cortez and M. Sameiro Carvalho
  Journal: Submitted to a Journal (re-submitted after first revision process)
  RO: RO3

1.5 Document Structure

This thesis is divided into six chapters. With the exception of the first and the last chapters, the remaining chapters were structured using the research papers written during this PhD program. The six chapters are structured as follows:

- **Chapter 1 - Introduction**
  This first chapter presents the main motivations behind this thesis, the research problem and opportunities identified, the outlined objectives, the research methodology adopted and also the structure of chapters regarding this document.

- **Chapter 2 - Background**
  The second chapter is related to important background concepts related to this thesis, namely Industry 4.0, Decision Support Systems and Business Analytics, Business Analytics and Industry 4.0 and generic concepts related to SC and Logistics.

- **Chapter 3 - State of Art**
  Chapter 3 presents a critical literature review regarding main topics surrounding this PhD thesis. It covers a systematic literature review about dimensioning safety stock under uncertainties and risks in the procurement process. Several literature gaps and research opportunities were highlighted, which provides a road map to guide future research regarding this topic. The work developed related this chapter was presented in the following journal article:

  - Júlio Barros, Paulo Cortez, M. Sameiro Carvalho, **A systematic literature review about dimensioning safety stock under uncertainties and risks in the procurement process**, Operations Research Perspectives, Elsevier, **8**:100192 (2021)
    DOI: https://doi.org/10.1016/j.orp.2021.100192

Note that, the development of the BDW is framed within the research project in which this thesis is as well inserted, and both of approaches proposed in Chapter 4 and 5 are used to validated the BDW.

- **Chapter 4 - Supervised learning of supply delay risk**
  This chapter focus on the first research problem and comprises the development of the proposed framework for predicting supplier delay risk. Moreover, comprises also the specification of the implemented BDW, in which aims to store, integrate and provide real data the proposed framework. The BDW was designed and implemented using the constructed data model, basing on the proposed logical and technological architectures. Focusing in the upstream SC, supply delays have a strong impact regarding the overall inventory management performance, often leading to SC disruptions. As such, identifying supply delay risk in a proactive fashion reveals to be valuable to organization towards improving the efficiency of inventory management process and consequently leads to cost saving. We address this research problem adopting the ML and BDA modeling techniques. Thereafter, we adopted a realistic and robust RW scheme in order to compare the predictive power, as well the supply chain inventory-related costs (namely, special freight and holding costs) of the six tested learning classification models. To the best of our knowledge, this work is the first in the SC management literature that combines the adopted modeling techniques in the context of supply risk identification. It should be essential to highlight that the implementation of BDW is not the main scope of this thesis. The proposed ML-based approaches use the data stored in the BDW and, on the other hand, are used to validate the BDW implementation. The work developed in this chapter resulted in the following journal articles:

  - Silva, N.; Barros, J.; Santos, M.Y.; Costa, C.; Cortez, P.; Carvalho, M.S.; Gonçalves, J.N.C. **Advancing Logistics 4.0 with the Implementation of a Big Data Warehouse: A Demonstration Case for the Automotive Industry**, Electronics **2021**, 10(18), 2221 DOI: https://doi.org/10.3390/electronics10182221
  - Júlio Barros, João N.C. Gonçalves, M. Sameiro Carvalho, Paulo Cortez, **A Machine Learning-based framework for predicting supply delay risk using Big Data technology**, submitted to a Journal

- **Chapter 5 - Supervised learning of estimating supply lead-time**
  Chapter 5 tackles the second research problem. It covers the development of the artifact addressed regarding the safety stock estimations. It is well-kwnow that enhancing the estimation of demand and LT leads to the improvement of safety stock estimations. Therefore, in this work, we focus on the enhancement of the LT estimation. LT has been given scarce attention compared to the demand uncertainty even though it constitutes a core parameter that varies and affects the SC performance and inventory parameters. In that sense, we introduce a IDSS that aims to enhance the LT uncertainty prediction supported by a BDW described in the previous chapter. Moreover, it

also provides a systematic approach to determine safety stock (using a well-known method used in the literature) minimizing the holding inventory costs while attending to a certain Service Level (SL). We evaluated empirically this real-world case study regarding the predictive power of five tested learning regression models and inventory holding costs from the selected model. To the best of our knowledge, this work is the first one on the SC management literature that combines ML and BDA to predict supply LT for the Material Requirements Planning (MRP) environment in the multinational automotive electronics industry (Bosch AE/P). The work developed in this chapter resulted in the following journal article:

  - Júlio Barros, João N.C. Gonçalves, M. Sameiro Carvalho, Paulo Cortez, **A machine learning strategy for estimating supply lead times towards improved safety stock dimensioning**, submitted to a Journal

- **Chapter 6 - Conclusion**
  This last chapter concludes this document, summarizing the scientific contributions and exposing the limitations of the developed work. In addition, suggestions for future work are also presented.

2 Background

Summary: This chapter provides an overview of fundamental concepts related to the content of this doctoral thesis. It starts introducing the Industry 4.0, focusing on providing an overview of the industrial revolution, followed by the definition of Industrial Internet of Things (I-IoT) and Industrial Cyber-Physical Systems (I-CPS). The second section comprises concepts related to Decision Support Systems (DSS) and Business Analytics (BA), followed by BA and Industry 4.0. Lastly, this chapter ends with an overview of Supply Chain (SC) and Logistics and Logistics processes at Bosch Automotive Electronics, Portugal (AE/P).

Chapter Table of Contents:

2.1 Industry 4.0 . . . 15
2.1.1 Overview of the industrial revolution . . . 15
2.1.2 Industrial Internet of Things . . . 15
2.1.3 Industrial Cyber-Physical Systems . . . 17
2.2 Decision Support Systems and Business Analytics . . . 18
2.2.1 Decision Support Systems (DSS) . . . 18
2.2.2 Machine Learning and Predictive Analytics . . . 20
2.3 Business Analytics and Industry 4.0 . . . 22
2.4 Supply Chain and Logistics . . . 23
2.4.1 Procurement . . . 26
2.4.2 Production . . . 27
2.4.3 Distribution . . . 28
2.4.4 Customer Interface . . . 29

2.5 Logistics Processes at Bosch Automotive Electronics 30
2.5.1 Demand planning 33
2.5.2 Production planning 33
2.5.3 Procurement planning 34

2.1 Industry 4.0

The industrial Hannover Fair (Germany) in 2011 was the place where the term "Industry 4.0" was firstly used (Boyes et al., 2018; Drath & Horch, 2014; Y. Liao et al., 2017; Wollschlaeger et al., 2017) and was raised numerous discussions and the major question was: is it a hit or hype? This term is referred to as the application of generic concepts of Cyber-Physical Systems (CPS) to industrial production systems (Drath & Horch, 2014). Industry 4.0 represents the introduction of the information technologies into the industry so that to achieve a higher level of operational efficiency, productivity, and automatization (Drath & Horch, 2014; H. Xu et al., 2018; L. Yang, 2017). It brings a set of disruptive technologies that are transforming industrial production, business models and business processes, such as, autonomous robots, simulation, cybersecurity, cloud computing, augmented reality, artificial intelligence, big data and analytics, and other technologies. The design principles of industry 4.0 are interoperability, virtualization, decentralization, realtime capability, service orientation, and modularity (L. Yang, 2017). Two key concepts emerged with industry 4.0: I-IoT and I-CPS. I-IoT and I-CPS are extensions of the traditional Internet of Things (IoT) and CPS. IoT is directed to a consumerbased system and on the other hand, I-IoT consists of the interconnection of intelligent industrial devices in order to improve the operational efficiency and productivity of the industrial system. In a similar way, the CPS are directed to critical infrastructures or consumer application and I-CPS consists to support the manufacturing and industrial production applications (H. Xu et al., 2018).

2.1.1 Overview of the industrial revolution

Figure 4 represents the evolution of the industry. The first industrial revolution started at the end of the 18th century and was the mechanizations of production using water and steam power. The second industrial revolution started at the beginning of the 20th century and was the introduction of mass production powered by electricity, and combustion engines and the introduction of assembly lines. The third industrial revolution began in the 1970s and was the digital revolution. This revolution represented the introduction of electronics, IT technologies and industrial robotics for advanced automation of the production process (Drath & Horch, 2014; Y. Liao et al., 2017; Wollschlaeger et al., 2017; L. Yang, 2017). The fourth industrial revolution represents the digitalization of the industry. Consists of integrating CPS with production, logistics, and services in the current industrial practices (Drath & Horch, 2014; Wollschlaeger et al., 2017; L. Yang, 2017).

2.1.2 Industrial Internet of Things

There are numerous definitions for the internet of things in the literature and the three more relevant definitions are (Boyes et al., 2018):

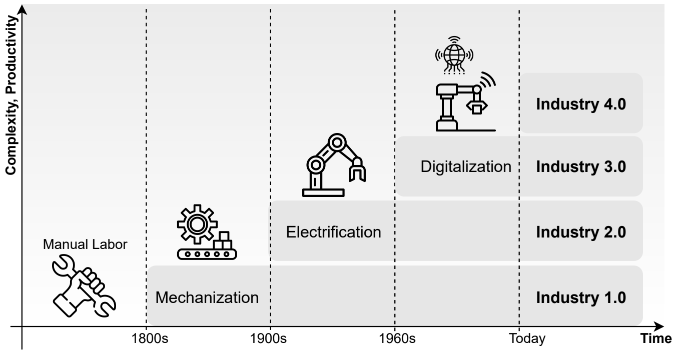


Figure 4: Overview of industrial revolutions adapted from (Drath & Horch, 2014)

- *“A definition for the IoT would be a group of infrastructures, interconnecting connected objects and allowing their management, data mining and the access to data they generate where connected are sensor(s) and/or actuator(s) carrying out a specific function that is able to communicate with other equipment”*;

- *“The terms Internet of Things and IoT refer broadly to the extension of network connectivity and computing capability to objects, devices, sensors, and items not ordinarily considered to be computers. These smart objects require minimal human intervention to generate, exchange, and consume data; they often feature connectivity to remote data collection, analysis, and management capabilities”*;

- *“The IoT represents a scenario in which every object or thing is embedded with a sensor and is capable of automatically communicating its state with other objects and automated systems within the environment. Each object represents a node in a virtual network, continuously transmitting a large volume of data about itself and its surroundings...”*.

Basing in these definitions of IoT, I-IoT can be defined as a global network infrastructure that allows interconnection of industrial devices and equipment’s through sensory, communication, networking, and information processing technologies, so that to share information between them and coordinate decisions (Al-Fuqaha et al., 2015; Boyes et al., 2018; H. Xu et al., 2018; L. D. Xu et al., 2014), this is, I-IoT consists to use IoT technologies in industrial environments in order to interconnect industrial assets (smart objects and cyber-physical assets) (Boyes et al., 2018; H. Xu et al., 2018). And this network infrastructure can be used to monitor and control physical objects in CPS (H. Xu et al., 2018). The generic I-IoT architecture consists of three layers, as depicts in Figure 5. These three layers are: application layer, communication/network layer and physical layer (Al-Fuqaha et al., 2015; H. Xu et al., 2018; L. D. Xu et al., 2014). The application layer refers to different industrial applications, such as smart factories, smart plants, smart supply chains, and other applications. The main responsibility of this layer is to link the gap between user and applications. This layer provides a timely monitoring, accurate control, and efficient management through numerous sensors and actuators in those smart industrial applications (Al-Fuqaha et al., 2015;

H. Xu et al., 2018). The communication/network layer consists to connect all *"things"* through numerous communication networks and technologies, such as Wireless Sensor and Actuator Network (WSAN), 5G, Wireless Personal Area Networks (WPAN), Machine-to-Machine (M2M), Software-Defined Networking (SDN) in order to share information's between them, this is, the communication layer provides networking support and data transfer for communication between *"things"* (H. Xu et al., 2018; L. D. Xu et al., 2014). Finally, the physical layer is composed by smart objects and cyber-physical assets, such as sensors, actuators, Radio Frequency-based Identification (RFID), manufacturing equipment's, and other industrial objects (Al-Fuqaha et al., 2015; H. Xu et al., 2018; L. D. Xu et al., 2014). These smart objects and cyber-physical assets are responsible for acquiring and processing information (Al-Fuqaha et al., 2015; L. D. Xu et al., 2014).

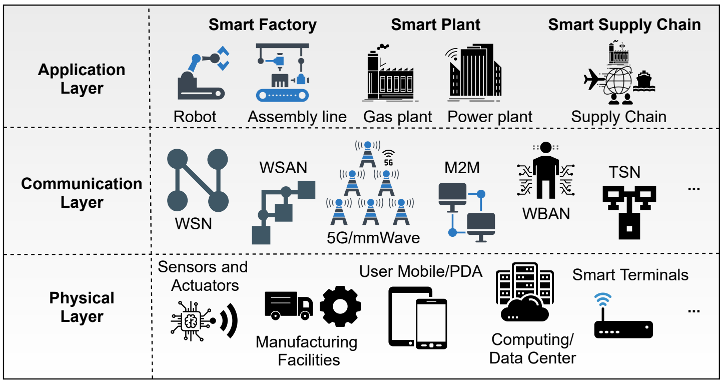


Figure 5: I-IoT Architecture adapted from (H. Xu et al., 2018)

2.1.3 Industrial Cyber-Physical Systems

In 2006, Helen Gill proposes the term CPS in the National Science Foundation (NSF) CPS workshop and after those numerous definitions of CPS were proposed in the literature (Alguliyev et al., 2018). Some example of those definitions are:

- *"CPS is a system that can effectively integrate cyber and physical components using the modern sensor, computing and network technologies" (Alguliyev et al., 2018)*;

- *"CPS is the integration of computing and physical processes. They include embedded computers, network monitors, and controllers, usually with feedback, where physical processes affect computations and vice versa" (Alguliyev et al., 2018; Colombo et al., 2017)*;

- *"CPS are systems of collaborating computational entities which are in intensive connection with the surrounding physical world and its on-going processes, providing services available on the internet" (Monostori et al., 2016)*;

- *"A system comprising a set of interacting physical and digital components, which may be centralized or distributed, that provides a combination of sensing, control, computation and networking functions, to influence outcomes in the real world through physical processes" (Boyes et al., 2018).*

I-CPS is defined as a vertical industrial system based on cyber and physical systems (H. Xu et al., 2018), as depicts in Figure 6. Every real physical object has at least one cyber representation, and each cyber system can be associated with a physical representation (Colombo et al., 2017). I-CPS provides productive and efficient manufacturing and automation, and enable the monitoring and control of industrial physical processes (Colombo et al., 2017; H. Xu et al., 2018). The I-IoT represents the integration of communication/network layer of the I-CPS (H. Xu et al., 2018(H. Xu et al., 2018).

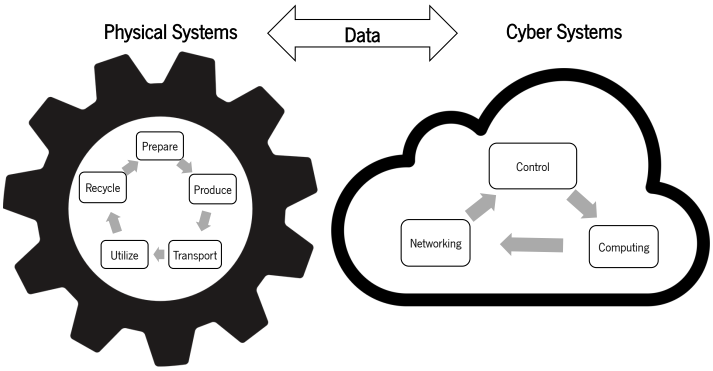


Figure 6: I-CPS Architecture adapted from (H. Xu et al., 2018)

2.2 Decision Support Systems and Business Analytics

This section presents concepts related to Decision Support Systems, Machine Learning and Predictive Analytics.

2.2.1 Decision Support Systems (DSS)

DSS is the area of Information System (IS) discipline that consists to support and improve the decision making (Arnott & Pervan, 2016). This concept was proposed in the early 1970s by Scott-Morton and was defined as *"interactive computer-based system, which helps decision makers utilize data and models to solve unstructured problems"*. Keen and Scott-Morton provided another DSS definition: *"Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision-makers who* *deal with semi-structured problems"* (Turban et al., 2011). The term Intelligent Decision Support System (IDSS) was termed at the early 1980s and aimed to embed the artificial intelligence and expert system tool into DSS (Arnott & Pervan, 2016; Kaklauskas, 2015). It consisted of an interactive tool for decision making for the well-structured decision, planning situation that used expert system techniques and specific decision models (Arnott & Pervan, 2016).

After a few years, in the early 1980s emerged the concept of Executive Information Systems (EIS) from the DSS and this new concept expanded the computerized support to top-level managers and executives through the introduction of the Online Analytical Processing (OLAP) tool that enables users to analyze multidimensional data interactively from multiple perspectives (Turban et al., 2011).

The term Business Intelligence (BI) was proposed by Howard Dresner in 1989 and it gained a widespread attraction until the early 2000s (Arnott & Pervan, 2016; H. Chen et al., 2012; Turban et al., 2011). The change from executive information systems to BI is warranted through the introduction of dimensional modeling and data warehousing concepts (Arnott & Pervan, 2016; Turban et al., 2011). The BI concept consists to combine architectures, tools, databases, analytical tools, applications, and methodologies (Turban et al., 2011).

Finally, the concept of BA has also risen, and Thomas H. Davenport popularized this new concept through his widely real professional article in the Harvard Business Review in 20006 (Arnott & Pervan, 2016; H. Chen et al., 2012). It emerged by the junction of the BI with a set of new capabilities, such as optimization, forecasting, predictive modeling, and statistical analysis. Davenport and Harris defined BA as *"the extensive use of data, statistical and quantitative analysis, exploratory and predictive models, and factbased management to drive decision and actions"* (Arnott & Pervan, 2016). Because the definition provides by Davenport is quite similar to the definition of BI and a wide range of practitioners do not distinguish the differences between BA and BI, H. Chen et al. (2012) suggested regarding the Davenport's definition that the term BA represents the key analytical component in BI. Consequently, is proposed Business Intelligence and Analytics (BI-A) as a unified concept (H. Chen et al., 2012). All these terms, IDSS, BI, BA originated from the evolution of the DSS are illustrated in Figure 7.

There are four levels of analytics: Descriptive, Diagnostic, Predictive and Prescriptive analytics. The value chain model of analytics developed by the Gartner Group represents a better visualization of these levels, as depicted in Figure 8:

- **Descriptive Analytics** - consists to answer the question such as *"What happened?"*, this is, consists to know what is happening in the organization (Delen & Demirkan, 2013; Koch et al., 2015; Sharda et al., 2015; G. Wang et al., 2016);

- **Diagnostic Analytics** - is related to the question *"Why did it happen?"*, this is, consist to identify the cause of the problem (Koch et al., 2015; Sharda et al., 2015);

- **Predictive Analytics** - attempts to answer the question *"What will happen?"*, this is, aims to determine what is likely to happen in the future (Delen & Demirkan, 2013; Koch et al., 2015;

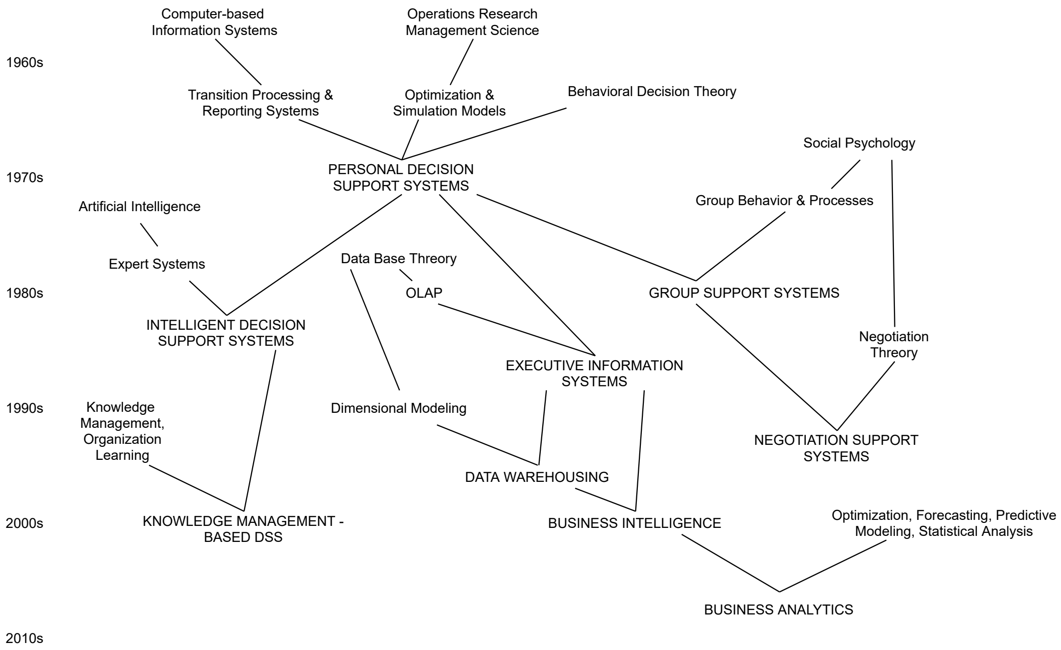

Computer-based Information Systems
Operations Research Management Science
1960s
Transition Processing & Reporting Systems
Optimization & Simulation Models
Behavioral Decision Theory
Social Psychology
1970s
PERSONAL DECISION SUPPORT SYSTEMS
Group Behavior & Processes
Artificial Intelligence
Expert Systems
Data Base Threory
1980s
OLAP
GROUP SUPPORT SYSTEMS
INTELLIGENT DECISION SUPPORT SYSTEMS
Negotiation Threory
EXECUTIVE INFORMATION SYSTEMS
1990s
Knowledge Management, Organization Learning
Dimensional Modeling
NEGOTIATION SUPPORT SYSTEMS
DATA WAREHOUSING
2000s
KNOWLEDGE MANAGEMENT - BASED DSS
BUSINESS INTELLIGENCE
Optimization, Forecasting, Predictive Modeling, Statistical Analysis
BUSINESS ANALYTICS
2010s

Figure 7: Overview of the evolution of DSS adapted from (Arnott & Pervan, 2016)

Sharda et al., 2015). It uses mathematical algorithms and programming to find explanatory and predictive patterns within data (Delen & Demirkan, 2013; G. Wang et al., 2016);

- **Prescriptive Analytics** - aims to answer the question *"How can we made it happen?"*, this is, aims to recognize the likely forecast and decision- make to achieve the best performance (Delen & Demirkan, 2013; Koch et al., 2015; Sharda et al., 2015; G. Wang et al., 2016). Prescriptive analytics include multi-criteria decision-making, optimization, and simulation (Delen & Demirkan, 2013; G. Wang et al., 2016).

The Predictive and Prescriptive analytics are crucial elements in helping companies to make an effective decision regarding the strategic direction of the organization. For problems such as the changes in organizational culture, sourcing decisions, supply chain configuration, and design and development of products or service, these two levels of analytics can be applied (G. Wang et al., 2016).

2.2.2 Machine Learning and Predictive Analytics

Machine Learning (ML) is considered a branch of Artificial Intelligence (AI) technologies that focus in the design and development of algorithms that allow computers learn based on historical data (Turban et al., 2011). ML uses computers programs to automatically learn complex patterns and make intelligent decisions based on data (Domingos, 2012; Han et al., 2012). ML has three different categories: Supervised learning, Unsupervised learning and Reinforcement learning (Han et al., 2012; Turban et al., 2011).

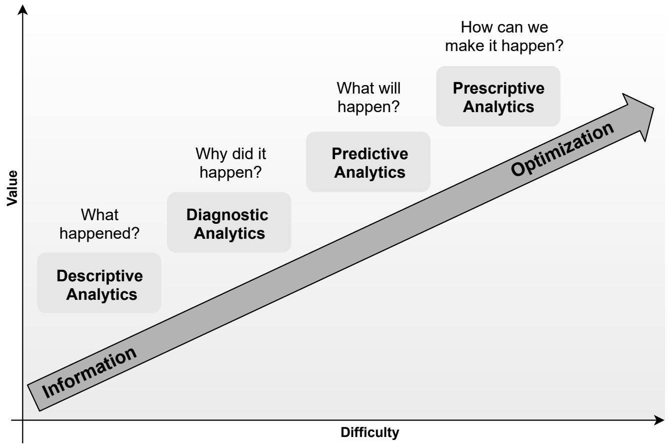


Figure 8: Gartner's Value Chain Model of Analytics adapted from (Koch et al., 2015)

- **Supervised learning** - this process consists to induce knowledge from a set of observations that include known outcomes, this is, algorithms learn from labeled data (Han et al., 2012; Turban et al., 2011). The supervised learning algorithms include classification and regression (Russel & Norving, 2010; Turban et al., 2011);

- **Unsupervised learning** - consists of discovery knowledge from a set of data without explicit outcomes, this is, the algorithms learn from unlabelled data (Turban et al., 2011). Typically, is used clustering to discover classes within the data (Han et al., 2012). The unsupervised learning algorithms include clustering segmentation and association (Turban et al., 2011);

- **Reinforcement learning** - in this process the algorithms learn from a series of reinforcements – rewards - good-result information or punishments - bad result information (Russel & Norving, 2010). This category differs from supervised learning because there is no historical data to learn, and from unsupervised learning because there is no natural grouping of things. This type of learning is applied to control the flight of helicopters, in autonomous search robots and other situations. The reinforcement learning algorithms include Q-Learning, Adaptive Heuristic Critic (AHR), State-Action-Reward State-Action (SARSA), Genetic Algorithms and Gradient Descent (Turban et al., 2011).

Figure 9 represents the categories of ML and exemplary methods for each category.

Predictive analytics consists of diverse techniques that predict the future based on historical and current data. The core of Predictive analytics aims to uncover patterns and determine relationships in data (Gandomi & Haider, 2015). It includes statistical models and other empirical methods that aim to create empirical prediction and methods for assessing the quality of those predictions in practice, this is, predictive power. Predictive power or predictive accuracy represents the ability of models to generate accurate predictions of new observation (Shmueli & Koppius, 2011).

Predictive analytics techniques can be divided into two groups (Gandomi & Haider, 2015):

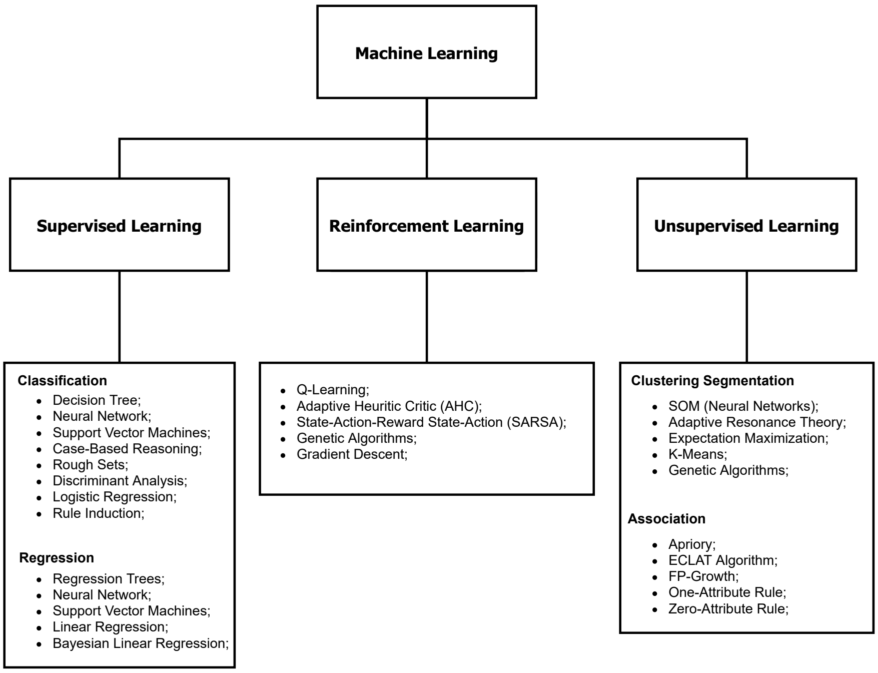


Figure 9: ML categories and respective methods adapted from (Turban et al., 2011)

A. techniques that consist to discover the historical patterns from the outcome variables and extrapolate them to the future, such as the moving averages;

B. and, techniques that capture the interdependency between the outcome variables and explanatory variables and exploit them to predict the future, such as the linear regression.

2.3 Business Analytics and Industry 4.0

Industry 4.0 represents the introduction of the information technologies into the industry to achieve a higher level of operational efficiency, productivity, and automatization (Drath & Horch, 2014; H. Xu et al., 2018; L. Yang, 2017). The I-IoT and I-CPS are the two key concepts that emerge with industry 4.0, transforming traditional factories into smart factories (H. Xu et al., 2018). These factories of the future or smart factories generate a massive amount of industrial data from a wide range of sources, such as the Enterprise Resource Planning (ERP) systems, distributed manufacturing environments, orders, and shipment logistics, customer buying patterns, product lifecycle operations, and technology-driven data sources, such as Global Positioning Systems (GPS), RFID tracking, and others sources (Božič & Dimovski, 2019; Govindan et al., 2018; Trkman et al., 2010; Waller & Fawcett, 2013; G. Wang et al., 2016). The generated data can be converted to valuable insights for the company through data analysis and integration. In this context, Business analytics and big data analytics have come up with tools and techniques to improve the decision-making process and create business value and competitive advantages to companies (Božič & Dimovski, 2019; Waller & Fawcett, 2013; G. Wang et al., 2016). BA has been increasingly considered to become an integral part of the organization business process and provides several benefits to companies. Such benefits are the increasing of revenues, increase of customer satisfaction, increase of product quality, better resource planning, better insights on customer needs, optimized supply chain, better demand forecast, lower cost base (cost cutting), better compliance with regulations, and others benefits, this is, the main benefits are the improvement of the operational efficiency and the empowerment of the organizational (Božič & Dimovski, 2019; Lueth et al., 2016; Trkman et al., 2010). BA can be used to solve generical problems from different areas of the industry, such as manufacturing, and logistics and supply chain, in order to enhance organizational performance. In manufacturing/operation, BA can be applied for predictive maintenance of equipment, machinery and assets (e.g., rescheduling the maintenance plan so that to act before the equipment failure according to historical and real-time machine performance analysis), decision-support system for industrial processes (e.g., using data from operations to automate purchase order or production scheduling decisions), manufacturing network optimization (e.g., correlating and optimizing performance across multiple plants), and optimizing individual machine parameters for smooth operations and optimal quality (e.g., correlating cause and effect of parameters such as machine speed). In the logistics/supply chain it can be applied for monitoring of moving assets (e.g., goods in transit), cross-supplier supply chain optimization (e.g., analysing warehouse stock level and real-time supply data to forecast shortages, reduce overall inventory levels and bring efficiency to the supply chain), fleet management (e.g., analysing transportation data and fuel consumption to optimize the distribution network), and strategic supplier management (e.g., continuously analyse quality metrics of individual suppliers) (Lueth et al., 2016). There are several examples of practical application of BA in industry, such as the HPE – predictive maintenance of wind turbines from the Hewlett Packard Enterprise that use Predictive/ML techniques on data collected from turbines to predict when the wind turbine needs maintenance. Another example is provided by the Comma Soft AG that uses a BA tool for reducing the complexity-driven cost that consists to use optimizing available product variants through the elimination of rarely chosen product variants and very expensive product options that led to millions in saving of costs (Lueth et al., 2016). Moreover, DHL also uses BA for combining the external operational and macroeconomic data to improve the operation efficiency of their supply chain (G. Wang et al., 2016).

2.4 Supply Chain and Logistics

Often times, Supply Chain Management (SCM) is confused with Logistics Management (LM) and this situation led to the Council of Supply Chain Management Professional (CSCMP) to propose the official definitions for these two terms, as following (Council of Supply Chain Management Professional - CSCMP, 2013a):

- *"Supply chain management encompasses the planning and management of all activities involved in sourcing and Procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third-party service providers, and customers. In essence, supply chain management* *integrates supply and demand management within and across companies"*;

- *"Logistics management is that part of supply chain management that plans, implements, and controls the efficient, effective forward and reverses flow and storage of goods, services and related information between the point of origin and the point of consumption in order to meet customers' requirements"*.

SCM includes all activities of logistics management and manufacturing operations. Also, it coordinates the activities and processes of different areas, such as marketing, sales, product design, finance, and information technology. On the other hand, logistics management activities involve inbound and outbound transportation management, fleet management, warehouse, materials handling, order fulfillment, logistics network design, inventory management, supply/demand planning, and management of third-party logistics services providers. Others functions such as sourcing and Procurement, production planning and scheduling, packaging and assembly, and customer services are assigned to the logistics (Council of Supply Chain Management Professional - CSCMP, 2013a).

SCM includes five main processes, as depicts in Figure 10 (Council of Supply Chain Management Professional - CSCMP, 2013b):

- **Planning or Planning process** - is carried out over all remaining supply chain processes. It aims to create effective long and short-range strategies for the supply chain. The strategies for the integrated supply chain should be developed from the design of the supply chain network to the prediction of the customer demand;

- **Procurement or Buying process** - focusing on the purchasing of the required raw materials, components, and goods;

- **Production or Make process** - includes the manufacture, conversion, or assembly of materials into finished goods or parts for other products;

- **Distribution or Moving process** - this process aims to manage the logistic flow of goods through the supply chain. This flow of goods is ensured by the transportation companies, third-party logistics firms, and others;

- **Customer Interface or Demand process** - corresponds to all issues that are related to planning the interactions with customers, satisfying their needs, and fulfill their orders;

Logistics management is based on the following five activities (Council of Supply Chain Management Professional, 2010):

- **Planning** - the planning process aims to plan activities related to the operationality of the supply chain. It includes gathering customer requirements, collecting information, and balancing requirements and resources in order to determine planned capabilities and resource gaps;

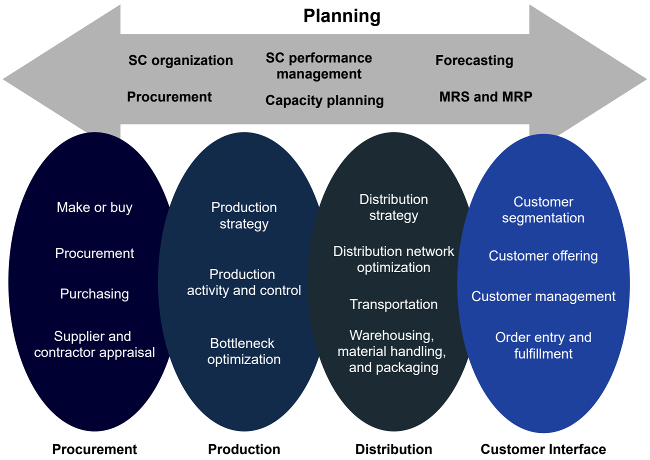


Figure 10: Supply Chain adapted from (Council of Supply Chain Management Professional - CSCMP, 2013b)

- **Source** - this process consists to determine the ordering or scheduling, and receipt of goods and services. It includes purchasing orders, scheduling deliveries, receiving, shipment validation and storage, and accepting supplier invoices;

- **Make** - this process specifies the activities related to the conversion of materials or the creation of the content for services. It is focusing on the conversion of materials, such as assembly, chemical processing, maintenance, repair, overhaul, recycling, refurbishment, remanufacturing, among others;

- **Delivery** - this process describes the activities related to the creation, maintenance and fulfillment of the customer order. Deliver process includes the receipt, validation, and creation of customer orders, scheduling order delivery, shipment and invoicing the customer; and;

- **Return** - this process determines the activities related to the reserve flow of goods back from the customer. It includes the identification of the need of return, the disposition decision making, the scheduling of the return, and the shipment and receipt of the returned goods.

Logistics consists of a set of fundamental supply chain processes that facilitates the fulfillment of the customer's demands. It aims to supply the right product or service, the right customer, at the right time, at the right place, at the right condition, in the right quantity and at the right cost as depicts in Figure 11 (Gibson et al., 2014).

The Right PRODUCT
To the Right CUSTOMER
At the Right TIME
To the Right PLACE
In the Right CONDICTION
In the Right QUANTITY
At the Right COST

Figure 11: Logistics Management adapted from (Council of Supply Chain Management Professional - CSCMP, 2013b)

As described previously, logistics management activities involve inbound and outbound transportation management, fleet management, warehouse, materials handling, order fulfillment, logistics network design, inventory management, and others. These activities can be grouped into processes, such as Procurement, production, distribution and customer interface. In the following sub-sections is described these processes.

2.4.1 Procurement

The terms purchasing and procurement are often used as the same concept, although they differ in scope. Purchasing is related to the actual buying of materials and the buying process activities. In the other hand, Procurement has a broader scope comparing with purchasing. It includes purchasing, warehousing, and all activities of receiving inbound materials (Lambert et al., 1998). Purchasing is the first step in procurement within a process-based supply chain.

Chopra and Meindl (2016) defines Procurement as *"the process of obtaining goods and services within a supply chain"*. Also, the Council of Supply Chain Management Professional - CSCMP (2013a) provides a definition for Procurement: *"the activities associated with acquiring products or services. The range of activities can vary widely between organization to include all of parts of the functions of Procurement planning, purchasing, inventory control, traffic, receiving, incoming inspection, and salvage operations"*.

Procurement and supply represent two of the key processes in the supply chain and can influence the success of the entire organization. They consist to ensure the sufficient supplies of raw materials at the right price, of the required quantity, in the right place and at the right time (Rushton et al., 2014). The typical Procurement cycle progress sequentially on the following steps (Rushton et al., 2014):

A. The identification of the need to procure a good or service;

B. Production of requisition document that needs to be approved and passed to the Procurement department;

C. A Request for Quotation (RFQ) is sent to a selection of suppliers;

D. Suppliers respond with prices and a period of negotiation may be entered into;

E. A supplier is selected, and Purchase Order (PO) is raised containing detailed records of the agreed price, delivery terms and place, and items or services to be provided;

F. The PO is signed and authorized by a manager. It is then sent to the supplier;

G. The goods or services are delivered and inspected;

H. The supplier sends an invoice;

I. The invoice is approved and paid or held pending resolution of any discrepancies found;

J. The Procurement department assesses the performance of the supplier based on quality, timelines, price and the completeness of the order. This is known as a post-contract review.

The procurement process includes activities such as the ”make or buy”decision process, purchasing, Procurement, and supplier and contractor appraisal. Figure 12 represents the Procurement cycle in a resumed way.

A. **Sourcing** - also known as Strategic procurement consists of a set of business processes that are required to purchase goods and services (Chopra & Meindl, 2016; N. R. Sanders, 2012). It includes processes such as formulize specification, selecting suppliers and contracting process (Carvalho et al., 2017; N. R. Sanders, 2012).

B. **Purchasing** - Purchasing or Operational procurement consists of the processes of buying goods and services (N. R. Sanders, 2012). It includes processes such as the ordering of material and services, monitoring and evaluation (Carvalho et al., 2017).

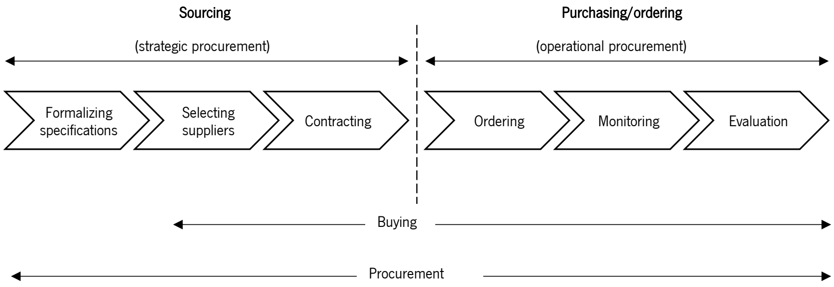


Figure 12: Steps of procurement process adapted from (Carvalho et al., 2017)

2.4.2 Production

The production process, also known as manufacturing logistics or Operations Management (OM), consists of all the activities that are involved in the production of goods and services (Rushton et al., 2014). Greasley (2009) provides a description of OM as *“Operations Management is about the management of the processes that produce or deliver goods and services. Not every organization will have a functional department called operations, but they will all undertake operations activities because every organization produces goods and/or services”* (Rushton et al., 2014; N. Sanders, 2014). It involves planning, organizing, coordinating, and controlling the resources needed to produce goods and services (N. Sanders, 2014). The OM is a system-based process, which the inputs of the system are converted into outputs through the transformation process, as depicts in Figure 13. In the context of logistics management, OM is the conversion process between the procurement and finished product storage, and distribution (Rushton et al., 2014). Is important to not confuse operations management with operational management.

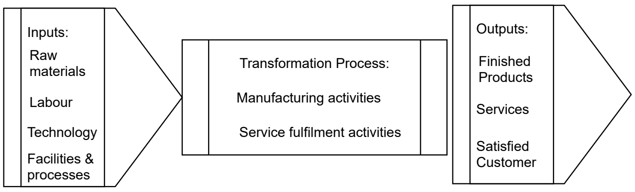


Figure 13: Operations Management process adapted from (Rushton et al., 2014)

The production, production planning, and production sequencing are increasingly important areas for logistics, due to the implications between these activities and the management of materials flow, and the management of the stocks (Carvalho et al., 2017; Lambert et al., 1998). Because the planning should start through the demand, the logistics have a key role to link the demand forecasting to the production, both at the level of planning and sequencing (Carvalho et al., 2017).

The production process includes activities such as production strategy, production activity, production control, and bottleneck optimization. Some of these activities or sub-activities of these activities are the critical ones and drives the success of the business operation, such as the production planning, Master Production Schedule (MPS), Material Requirements Planning (MRP), and Rough Cut Capacity Planning (RCCP) (N. Sanders, 2014).

2.4.3 Distribution

Distribution is related to all steps needed to move and store goods from the supplier to a customer in the supply chain. It includes activities such as distribution strategy, distribution network optimization, transportation, and warehousing, material handling, and packaging. This process effects directly the supply chain cost and customer value, thereby making a key driver of the company profits (Chopra & Meindl, 2016).

There are three different strategies that can be adopted for goods distribution, as depicts in Figure 14 (Chopra & Meindl, 2016; Ghiani et al., 2004; Simchi-Levi et al., 2000):

- **Direct shipment** - this strategy consists to ship goods directly from the manufacturer to the final customer without going through distribution centers;
- **Warehousing** - is the traditional approach used and aims to receive goods in warehouses and store them in tanks, pallet racks or on shelves. When one order is received, the items are retrieved, packed and shipped to the customer;
- **Cross-docking** - this strategy is also known as just-in-time distribution and consists of transhipments facilities where the incoming shipments are sorted, consolidated with other products and transferred directly to outgoing trailers without intermediate storage or order picking.

Transportation as an integral part of the distribution process is one of the most structured activities in logistics and is responsible for a large part of the logistics costs (Carvalho et al., 2017). Transportation

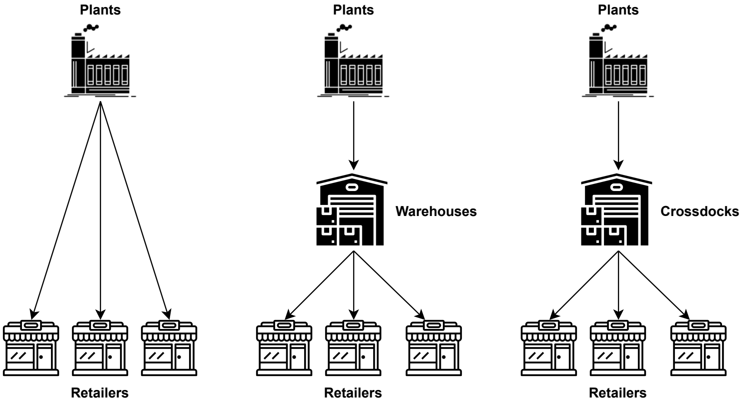


Figure 14: Distribution strategy process adapted from (Ghiani et al., 2004)

aims to provide the flow of inventory from the origin point in the supply chain to the destination and effects both responsiveness and efficiency (Chopra & Meindl, 2016; Goldsby et al., 2014; Lambert et al., 1998). Almost all the companies manage both inbound and outbound logistics, this is, to manage the Procurement of material and goods from suppliers and the distribution of materials and goods to customers. And to perform this, it is necessary to include the transportation (Goldsby et al., 2014). The selection of the transportation modes, the contracting of transport service providers (external to the company) and contractual management of these transport service providers are critical activities for logistics (Carvalho et al., 2017).

Another integral part of the distribution process is warehousing. Warehouses are one of the most crucial components of the logistics and consist to facilitate the movement of goods through the supply chain to the final customer (Lambert et al., 1998; Rushton et al., 2014). Warehouses need to be designed and operated regarding the specific requirements of the supply chain (Rushton et al., 2014).

Another activity very important in logistics, and also an integral part of the distribution process, is the industrial packaging in a sense of the protection of materials during transportation and storage of goods (Carvalho et al., 2017). The packaging use unit loads such as pallets, cage and box pallets, roll-cages, tote bins, dollies and Intermediate Bulk Container (IBC) (Rushton et al., 2014). Handling of materials is important and critical for several areas, including production planning, warehousing design, and its efficiency. There are various types of storage and handling equipment's for palletized and non-palletized goods, such as conveyors, forklift trucks, overhead cranes, retrieve (ASRS systems) (Carvalho et al., 2017; Lambert et al., 1998; Rushton et al., 2014).

2.4.4 Customer Interface

The customer interface can be considered an activity and either the principal output of the logistic system. As an activity or a set of activities, customer interface is related to the possibility to make available materials or services, this is, have the right product/material/service, to the right customer, in the right quantity, at right place, at the right time, in the right condition, at the right cost or the minimal cost.

Customer interface can be regarded as the availability of products/materials/services, but also as the process of interaction with the customer, in order to influence them to place an order. It can be also understood as the service levels (several dimensions and variables) that companies provide to their internal and external clients (Carvalho et al., 2017).

Companies measure the quality of their products or services through internal quality assurance, external customer satisfaction and customer value. Internal quality measures the number of defeats, customer satisfaction measure the satisfaction of clients with the company products and the impression of their service and customer value promote a broader view regarding the offering of the company and customers. This aims to go deeper and answer the following question: *"why customers purchase and continue to purchase?"*, *"what are the customer preferences and needs?"*, *"how can they be satisfied?"* and *"which customers may incur losses"*. Customer value is important to determine which type of supply chain is required to serve the customer and what services are required to retain customers (Simchi-Levi et al., 2000).

The customer interface process includes activities such as customer segmentation, customer offering, customer management, order entry, fulfillment and reverse logistics. Reverse logistics consists of moving of materials in the opposite direction, this is, from the end customer to suppliers (Carvalho et al., 2017; Rushton et al., 2014). Generally, the reverse logistics is linked to the product recall for quality or safety reasons, to return of unwanted goods and used packaging or products for recycling or disposal (Rushton et al., 2014). Companies have been developing more flexibles return policies in order to reduce the risk of customers, increase their satisfaction and consequently increase sales.

2.5 Logistics Processes at Bosch Automotive Electronics

The Logistics Planning Management (LPM) processes at Bosch aim at establishing the right product, in the right quantity, in the right condition, in the right place, at the right time, and at the right cost (minimal cost), in order to ensure the accomplishment of the Original Equipment Manufacturer (OEM) requirements. Figure 15 illustrates the process chain map that is common to all Bosch plants of Business sector Mobility Solutions (BBM).

The processes featured in blue in Figure 15 illustrates the LPM processes at Bosch. The main concepts are described in the next Sections (see Sections 2.5.1, 2.5.2 and 2.5.3).

In logistics, material and goods flow goes from suppliers to customers, and logistics planning scope goes from customer to suppliers, as depicted in Figure 16.

Logistics management adopted at Bosch stands on the Supply Chain Operation Reference (SCOR) model (see Figure 17), which is composed of the following six primary management processes:

- **Plan** - this process consists of developing plans to operate the SC. It includes collecting requirements, gathering information on available resources, balancing requirements and resources in

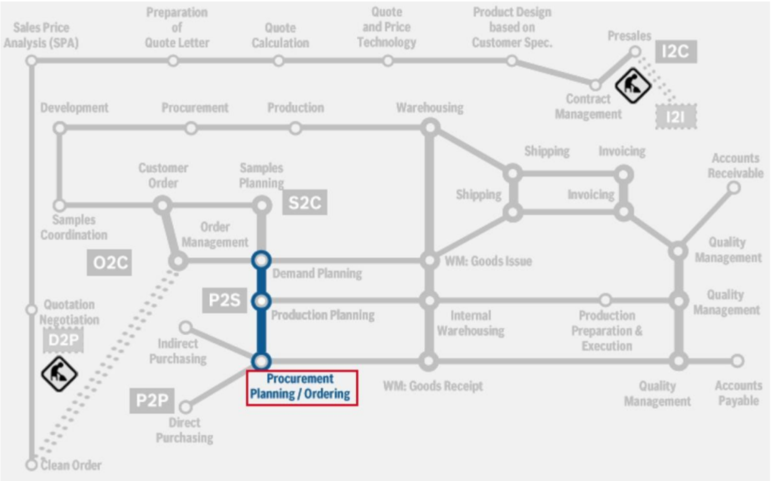


Figure 15: Process chain map of BBM.

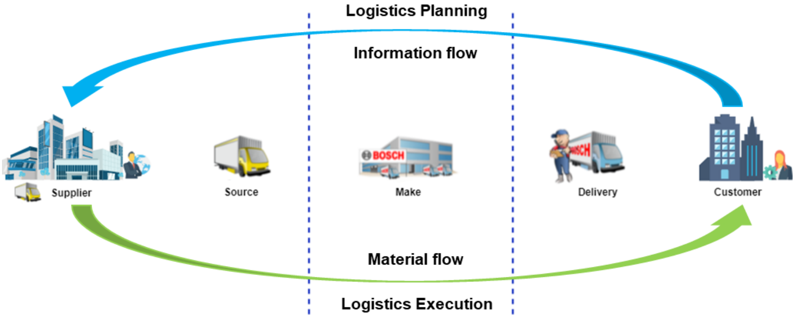


Figure 16: Logistics management value chain.

order to determine planned capabilities and gaps in demand or resources, and finally identifying corrective actions for these gaps;

- **Source** - this process aims to describe the ordering and receipt of goods and services, i.e., seeks to produce goods and services to meet planned or actual demand. The source processes include purchase order issuance, receiving, validation and storage of goods and accepting the invoice from suppliers;

- **Make** - this process consists of transforming products into a finished state in order to meet planned or actual demand. It describes all activities associated with the conversion of materials or creating of the content for services, which includes assembly, chemical processing, maintenance, repair, overhaul, recycling, refurbishment, remanufacturing and other common names for material conversion processes;

- **Deliver** - this process aims to describe the activities associated with creation, maintenance and fulfilment of customer orders. It includes the receipt, validation and performing customer orders, scheduling order delivery, pick, pack and shipment;

- **Return** - this process describes all activities related to the reverse flow of goods. It includes the identification of the need to return, disposition decision making, scheduling of the return, and shipment and receipt of the returned goods;

- **Enable** - this process describes all activities related to supply chain management. It includes the management of business rules, performance management, data management, resource management, facilities management, regulatory compliance management and risk management.

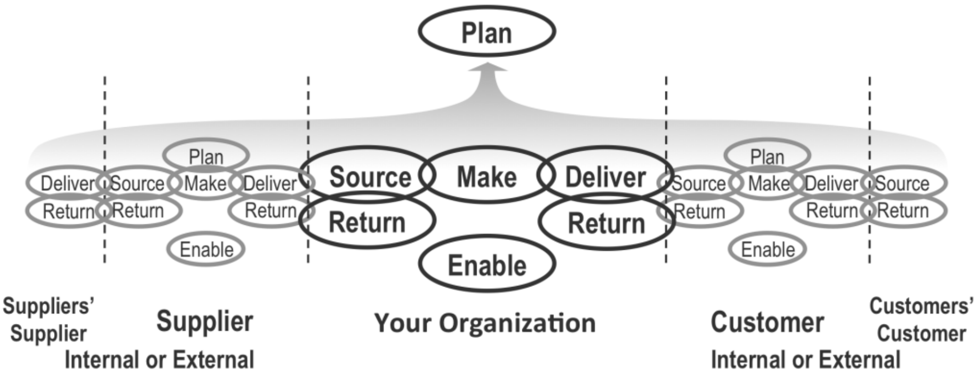
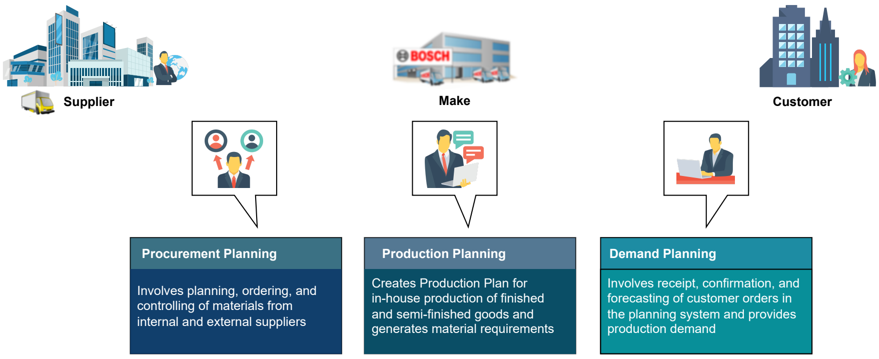


Figure 17: SCOR model adopted at Bosch.

The Bosch Production System (BPS) follows the push production philosophy, i.e., the production orders are carried out based on forecast demands and inventory information. The logistics management starts with the demand planning process (Delivery activity in Figure 16), which aims to plan and estimate the forecast of final product shipment to the OEM. After this first process, the production planning process (Make activity) starts to plan the internal production required to fulfil the OEM requirements. And finally, the procurement planning process (Source activity) seeks to identify and plan all the materials to manufacture the products ordered by the OEM. Figure 18 represents an overview of the three processes of LPM at Bosch.



Procurement Planning

Involves planning, ordering, and controlling of materials from internal and external suppliers

Production Planning

Creates Production Plan for in-house production of finished and semi-finished goods and generates material requirements

Demand Planning

Involves receipt, confirmation, and forecasting of customer orders in the planning system and provides production demand

Figure 18: Overview of procurement, production and demand planning processes.

2.5.1 Demand planning

Demand planning process consists of receiving, confirming and forecasting the OEM orders using the planning system, and afterwards providing the production demand information. The main functions of this process are:

- **Planning the OEM orders** - this process is associated with the process of examining and processing the orders from OEM, reviewing and confirming customer demand;
- **Forecast planning** - this process aims to plan personal and machine capacity, making feasible the identification of possible bottlenecks and provide long-term demand forecasts to Bosch suppliers;
- **International Production Network (IPN) planning** - this process is related to the organization of BBM divisions. All plants of an IPN produced finished goods of the same product subclass. IPN is composed of a principal (lead) plant and several sister plants, as depicted in Figure 19. The lead plants develop consolidated planning at least once a month for assigned finished goods.

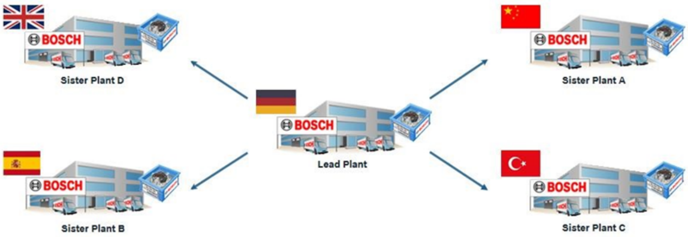


Figure 19: IPN planning process.

2.5.2 Production planning

The production planning process aims to plan the sequence of activities and required resources (manpower, machines, and materials) to produce a certain quantity in order to satisfy OEM requirements. It involves the creation of an assembly or production plan for in-house production of finished and semi-finished goods at Bosch and generates material requirements. Moreover, it also allows using the production capability through adequate production sequence, with no idle time or over-use, so that to maintain the inventory at an optimum level. The production planning represents the bond link between demand and procurement planning, as depicted in Figure 18. Two main phases composed this process:

- **Pacemaker** - this phase represents a process where the levelled customer signals are fed into production, defining the production rate (aims to align the production/manufacture to customer demand) and production sequence. The pacemaker involves creating levelling pattern (Pull) and sequence plan (Push), performing the production capacity check and measuring levelling adherence;

- **Production planning and control** - The production planning is a process that involves performing MRP run, creating a sequence plan (Push), performing the capacity check and reviewing MRP.

2.5.3 Procurement planning

Procurement planning is a process that involves continuous improvement of projects with suppliers and within the value stream. This process plays a fundamental role in managing and controlling the raw material inventory levels. There are three types of the procurement:

- **Standard procurement** - consists of acquiring products or services from suppliers;
- **Plant-to-plant** - represents the procurement of raw material or finished good in other Bosch plants;
- **Trading goods** - consists of purchasing materials from its suppliers by Bosch and then selling those materials to another company (is not used in any manufacturing operation).

The procurement planning aims to order materials required through production planning (Make activity). However, at Bosch, the procurement planning does not only consists of ordering the required materials, but also producing detailed planning and scheduling of the raw material deliveries from external and internal suppliers. Moreover, it also aims at meeting the raw material demands of the production lines and controlling the arrival of these materials. This process represents an interface between production planning (Make) and the suppliers (Source). Also, it involves several departments and activities, such as a raw material receipt, purchasing, inventory levels, warehouses, customer order planning, products planning, and logistics change management. It also supports the planning of some special situations, such as the ramp-up, ramp-down and reallocation.

The role of the procurement planning is to run the MRP for getting necessary raw materials and generating orders for these materials. MRP technique focuses for the inventory optimization.

3 State of Art

Summary: This chapter presents a Systematic Literature Review (SLR) regarding setting safety stock under uncertainties and risks in the procurement process. The goal is to identify gaps in the literature and research opportunities/trends used to guide the work inherent to this doctoral thesis (addressed in the next chapters).

Chapter Table of Contents:

3.1 Introduction 37
3.2 Theoretical background 39
3.2.1 Procurement: sourcing and purchasing 39
3.2.1.1 Sourcing 39
3.2.1.2 Purchasing 40
3.2.2 Supply chain risk and uncertainty 40
3.2.3 Uncertainty factors and risks in the procurement process 41
3.2.4 Dimensioning of safety stock (Traditional strategies) 43
3.3 Review methodology 46
3.3.1 Searching phase 46
3.3.2 Selecting phase 47
3.3.3 Analysing phase 47
3.4 Descriptive and co-occurrence analysis 48
3.4.1 Descriptive analysis 48
3.4.1.1 Year of publication 49
3.4.1.2 Venue of publication 49

3.4.1.3 Research problem 49
3.4.1.4 Authors influence and affiliations 50
3.4.1.5 Approach followed 52
3.4.1.6 Research method 52
3.4.2 Co-occurrence analysis 53
3.5 Literature analysis (Scientific contributions) 54
3.5.1 Safety stock dimensioning 54
3.5.1.1 Considering demand uncertainty 54
3.5.1.2 Considering lead time uncertainty 57
3.5.1.3 Considering yield uncertainty 59
3.5.1.4 Considering multiple uncertainties and risks 60
3.5.2 Safety stock management 62
3.5.2.1 Considering demand uncertainty 62
3.5.2.2 Considering lead time and yield uncertainty 65
3.5.2.3 Considering multiple uncertainties and risks 65
3.5.3 Safety stock allocation, positioning or placement 67
3.5.3.1 Considering demand and lead time uncertainty 68
3.5.3.2 Considering multiple uncertainties and risks 72
3.6 Literature gaps and research opportunities 74
3.7 Summary and directions for future research 77

A systematic literature review about dimensioning safety stock under uncertainties and risks in the procurement process

Júlio Barros[1] · **M. Sameiro Carvalho**[2] · **Paulo Cortez**[3] ·

Operations Research Perspectives, Elsevier, **8**:100192 (2021)
(https://doi.org/10.1016/j.orp.2021.100192)

Abstract This paper analyses literature contributions in the search for safety stock problem under uncertainties and risks in the procurement process, focusing on the dimensioning problem (determination of the safety stock level). We perform a SLR from 1995 to 2019 in relevant journals, covering 193 selected articles. These selected articles were classified into three safety stock main issues: safety stock dimensioning, safety stock management, and safety stock positioning, allocation or placement. The SLR analysis allowed the identification of literature gaps and research opportunities, thus providing a road map to guide future research on this topic.

Keywords Safety stocks, Inventory management, Procurement, Supply chain risk management, Uncertainty factors, Systematic literature review.

3.1 Introduction

The supply chain is a complex and unique network that integrates different business processes involved in fulfilling the customer needs, which includes planning, procurement, production, distribution and customer interface (Chopra & Meindl, 2016; Council of Supply Chain Management Professional - CSCMP, 2013a). All these are involved in the entire product life cycle, from procurement to manufacturing, distribution and customer service (Balfaqih et al., 2016). The importance of the supply chain management in business strategy, in attracting and retaining customers and markets, in the effectiveness of operation management and the profitability of companies results becomes a valuable way to ensure the competitive advantage and improving the organizational performance (Balfaqih et al., 2016; Carvalho et al., 2017; Trkman et al., 2010). Logistics plays an essential role in supply chain management and it is one of the crucial factors of the supply chain success. The logistics planning management processes aims at establishing the right product, in the right quantity, in the right condition, to the right place, at the right time, and at the right cost (i.e., minimal cost).

[1] ALGORITMI Research Centre, University of Minho, Guimarães 4800–058, Portugal.
[2] ALGORITMI Research Centre, Department of Production and Systems, University of Minho, Braga 4710–057, Portugal.
[3] ALGORITMI Research Centre, Department of Information Systems, University of Minho, Guimarães 4800–058, Portugal.

The supply chain management deals with a significant number of uncertainty factors that affect its performance. These uncertainty factors introduce a large number of random factors and events, affecting all dimensions of the supply chain activities, and also makes the risk and vulnerability a major challenge for organizations (Yu et al., 2015). Risks and uncertainty factors have a direct influence on both customer satisfaction levels and supply chain related costs. To deal with some of these factors, buffering techniques such as safety stock is included as the way for aiding the operational planning of manufacturing stages to cover both demand and supply uncertainties so that to provide the promised service level to the customers (C. A. Chang, 1985; Jung et al., 2008). Although a higher safety stock level represents a higher service level, it must be optimized in order to not increase the total costs of the supply chain (Jung et al., 2008).

Several authors have studied the safety stock research problem and proposed their inventory models considering different types of uncertainty and risks, using different approaches. The research problems related to safety stock involve typically issues such as dimensioning, management, and positioning, placement or allocation (Caridi & Cigolini, 2002a). Safety stock dimensioning consists of setting the appropriate safety stock level for each item. Safety stock management involves setting of both the safety stock levels and the time for replenishments. And, safety stock allocation, positioning or placement consists on setting safety stock levels and determine where to allocate them on supply chain structure. There are several terminologies in the literature for the same problem of safety stock placement. Safety stock placement, safety stock allocation and safety stock positioning represent the same problem (Graves & Willems, 2000; K. Kumar & Aouam, 2018b; H. Li & Jiang, 2012). In this Systematic Literature Review, we adopt the terminology safety stock placement to portray this problem.

Although the scope of this research is on safety stock dimensioning strategies, we extend it and consider all safety stock dimensions (this is, dimensioning, management and placement), since the dimensioning issue is present in each of these dimensions. Schmidt et al. (2012) argued that is very difficult to survey scientific publication related to safety stock dimensioning. Within our knowledge, there are only three surveys/reviews that cover totally or partially the safety stock problem. Caridi and Cigolini (2002a) analysed and classified safety stock damping methods for manufacturing systems by considering uncertainty factors. Schmidt et al. (2012) analysed mathematical methods for safety stock dimensioning and perform a simulation study to compare these methods regarding service and safety stock level. Finally, Eruguz et al. (2016) focused only on safety stock placement issue, more specifically on the guaranteed-service modelling approach.

A comprehensive SLR was made by analysing research papers from 1995 to 2019 of safety stock research efforts by considering uncertainty factors or risks, or even both, in the procurement process. The selected papers were filtered manually and reduced to 193 papers in this review and classified into three dimensions of safety stock problem: safety stock dimensioning, safety stock management, and safety stock placement, allocation or positioning. Furthermore, literature gaps were identified, allowing to disclose future research opportunities.

This paper is organized as follows. Firstly, Section 3.2 provides an overview of the main concepts related to procurement, supply chain risk and uncertainty, sources of uncertainty and risk in procurement processes and some traditional safety stock dimension strategies. Section 3.3 presents the review methodology followed for analysing the literature contributions. Section 3.4, we present a descriptive and co-occurrence analysis of selected papers. Then, in Section 3.5, the selected papers are categorized according to the research problem. Section 3.6 presents the literature gaps and research opportunities. Finally, we conclude this paper in Section 3.7.

3.2 Theoretical background

3.2.1 Procurement: sourcing and purchasing

The terms purchasing and procurement are often used as the same concept, although they differ in scope. Purchasing is related to the actual buying of materials and the buying process activities. On the other hand, procurement has a broader scope comparing with purchasing (Lambert et al., 1998; Monczka et al., 2010). It includes purchasing, warehousing, and all activities of receiving inbound materials (Lambert et al., 1998). Purchasing is the first step in procurement within a process-based supply chain.

Chopra and Meindl (2016) defined Procurement as *"the process of obtaining goods and services within a supply chain"*. Also, the Council of Supply Chain Management Professional - CSCMP (2013a) provides a definition for Procurement: *"the activities associated with acquiring products or services. The range of activities can vary widely between organization to include all of the parts of the functions of Procurement planning, purchasing, inventory control, traffic, receiving, incoming inspection, and salvage operations"*.

Procurement represents one of the key processes in the supply chain and can influence the success of the entire organization. It ensure the sufficient supplies of raw materials at the right price, of the required quantity, in the right place and at the right time (Rushton et al., 2014).

The procurement process includes activities such as the *"make or buy"* decision process, purchasing and appraisal of both supplier and contractor. Figure 20 represents the procurement cycle in a resumed way.

3.2.1.1 Sourcing

Sourcing, also known as strategic procurement consists of a set of business processes that are required to purchase goods and services (Chopra & Meindl, 2016; N. R. Sanders, 2012). It includes processes such as formalize specification, selecting suppliers and contracting process (Carvalho et al., 2017; N. R. Sanders, 2012).

- **Formalizing specifications** - in this process are defined the requirements of purchasing, as well as the *"make or buy"* decision (decision to make goods or provide a service rather than buying this goods/service) (Carvalho et al., 2017; Rushton et al., 2014). The first step of this process consists to define functional and technical specifications of items to be purchased (Carvalho et al., 2017);

- **Selecting suppliers** - this process consists of searching and identifying suppliers in the market (Carvalho et al., 2017; N. R. Sanders, 2012). Important decisions, such as the method of subcontracting to be adopted (e.g., partial or total subcontracting, payment in fixed-price or in refundable cost), the criteria for the preliminary qualification of potentials suppliers, the requisition and analysis of received proposals and selection of suppliers are necessary to be taken (Carvalho et al., 2017; Rushton et al., 2014);

- **Contracting** - in this process are defined the terms of the contract (e.g., delivery conditions and price, payments conditions, penalty clauses, and warranty conditions) and afterwards the signing of the contract (Carvalho et al., 2017).

3.2.1.2 Purchasing

Purchasing or operational procurement consists of the processes of buying goods and services (N. R. Sanders, 2012). For efficient purchasing is necessary to know the on-hand stock quantity so that to order the correct amount. An efficient purchasing requires inventory control management. Hence, safety stock as an extra inventory held to deal with uncertainties in demand and supply is used to plan future purchase quantities.

Purchasing includes processes such as the ordering of material and services, monitoring and evaluation (Carvalho et al., 2017).

- **Ordering** - this process consists to submit the purchasing order, but firstly is necessary to guarantee the definition of the contracting terms and consequently the signature of the contract;

- **Monitoring** - this process involves a set of different tasks related to the monitoring of submitted orders, such as visits to suppliers facilities, as well as negotiations related to changes regarding technical specifications, requisition of production plans and expected delivery date, verification of concordance of the delivered products with the agreed specifications, and lastly, the exchange of the commercial correspondences with customers;

- **Evaluation** - this process consists basically of the execution of complaints, activation of penalty clauses (when is applicable), and organization of documentation related to the project and supplier.

3.2.2 Supply chain risk and uncertainty

Often-times, the risk is confused with uncertainty, but these two terms are not the same (Colicchia & Strozzi, 2012; Sydow & Frenkel, 2013). Knight (1921) differentiate risk from uncertainty arguing that risk is

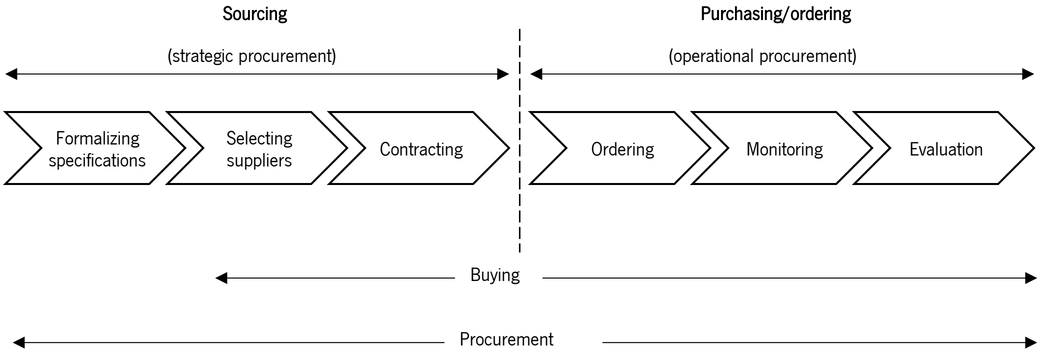


Figure 20: Steps of Procurement process adapted from (Carvalho et al., 2017)

something measurable while uncertainty is not quantifiable and unpredictable (with unknown outcomes). Manuj and Mentzer (2008) argues that risk is an expected outcome of an uncertain event, and Rao and Goldsby (2009) view risk as an event and uncertainty as possible outcomes.

There are several definitions in the literature regarding the risk in the supply chain context. But, do not exists a universal definition, although there have been several attempts (Baryannis, Dani, & Antoniou, 2019; W. Ho et al., 2015). Tables 1 and 2 presents some of the key definition of risks and Supply Chain Risks (SCR).

Table 1: Risk definitions

References	**Definitions**
March and Shapira (1987)	*"the variation in the distribution of possible SC outcomes, their likelihoods, and their subjective values."*
Royal Society (1992)	*"The probability that a particular adverse event occurs during a stated period of time, or results from a particular challenge. As a probability in the sense of statistical theory, risk obeys all the formal laws of combining probabilities."*
Mitchell (1995)	*"the probability of loss and the significance of that loss to the organization or individual."*
Harland et al. (2003)	*"a chance of danger, damage, loss, injury or any other undesired consequences."*

Table 2: SCR definitions

Reference	**Definitions**
Christopher (2003)	*"the identification of potential sources of risk and implementation of appropriate strategies through a coordinated approach among supply chain members, to reduce supply chain vulnerability."*
Jüttner et al. (2003)	*"any risks for the information, material and product flows from original supplier to the delivery of the final product for the end user."*
Zsidisin (2003)	*"supply chain risk is the potential occurrence of an inbound supply incident, which leads to the inability to meet customer demand."*
Peck (2006)	*"anything that (disrupts or impedes) the information, material or product flows from original suppliers to the delivery of the final product to the ultimate end-user."*
C. Tang (2006a)	*"the management of supply chain risks through coordination or collaboration among the supply chain partners so as to ensure profitability and continuity."*
W. Ho et al. (2015)	*"the likelihood and impact of unexpected macro and/or micro-level events or conditions that adversely influence any part of a supply chain, leading to operational, tactical or strategic level failures or irregularities."*

3.2.3 Uncertainty factors and risks in the procurement process

There are a variety of uncertainty factors and risks associated to the procurement process, such as uncertain lead time, demand fluctuations, variations of prices, uncertain yield, supplier delays and order crossover, as follows described.

- **Lead time uncertainty** - supply lead time represents the average of time between when the order is placed and when the product arrives (Chopra & Meindl, 2016; Disney et al., 2006). The uncertainty in supply lead time must be controlled properly in order to not increase the total cost and reduce customer service level (Hong, Lee, & Zhang, 2018). Besides that, the high variation of supply lead-time increases the difficulty in procurement planning (C. Ho et al., 2018), more properly to perform the MRP process. For an efficient production is necessary to estimate properly the procurement lead time and on-time delivery in order to prevent delays on deliveries that can lead to a shortage of inventory and consequently manufacturing disruption, increasing the total cost and revenue losses. Several strategies are used to cope with this type of uncertainty, such as safety stock, safety lead time and supplier backups. Safety stock is the most used strategy to increase the supply chain flexibility under both demand and supply uncertainty (Angkiriwang et al., 2014; Hong, Lee, & Zhang, 2018);

- **Demand uncertainty** - demand uncertainty includes factors such as errors in demand forecast, changes in customer orders and uncertainty about the product specification that the customers will order (Angkiriwang et al., 2014). Demand forecast consists to estimate the future Stock Keeping Units (SKU) in order to meet customer demands. The demand forecast is a complex task (C. Ho et al., 2018; Lambert et al., 1998) and when demand is not estimated accurately (forecast error) can lead to inventory short supply or surplus, low service level, rush orders, inefficient utilization of resources and bullwhip effect propagation along the supply chain (Chopra & Meindl, 2016; C. Ho et al., 2018; Nenni & Schiraldi, 2013). This type of uncertainty assumes an important role in the dimensioning of production lines, dimensioning of transportation modes, line assembly, distribution centres and cross-docking platforms (Carvalho et al., 2017) and also plays an important role as input for procurement planning (Nenni & Schiraldi, 2013). Component commonality, risk pooling, safety stock, safety lead time, flexible supply contracts, subcontracting/outsourcing and postponement are examples of strategies to cope with demand uncertainty (Angkiriwang et al., 2014; Hong, Lee, & Zhang, 2018);

- **Price uncertainty** - represents the fluctuations in the suppliers selling price of materials or raw materials due to the constant price fluctuation in the market or discount campaigns (Hong, Lee, & Zhang, 2018). Pricing must be considered as an important factor in the procurement process because it influences the logistics total cost, as well as the operational decisions (Choi et al., 2017). Flexible contract and price risk hedging are examples of strategies that can be used to deal with price uncertainty (Angkiriwang et al., 2014; Hong, Lee, & Zhang, 2018);

- **Yield uncertainty** - limited capability or defective products (quality issues) represent possible causes for yield uncertainty. There are two main approaches used to mitigate this type of uncertainty: supplier diversification (select multiples suppliers for unreliable supplier) and collaboration with suppliers (Hong, Lee, & Zhang, 2018). Another approach/strategy to cope with this uncertainty is capacity buffer (Angkiriwang et al., 2014; Hong, Lee, & Zhang, 2018);

- **Supplier delay** - on-time delivery is a standard objective of procurement and when is not properly estimate can lead to a shortage of inventory and consequently manufacturing disruption. Sometimes suppliers delays are caused by their quote delivery dates that cannot be achieved (Baily et al., 2015). Strategies such as supplier backups are the common strategies used to cope with this type of uncertainty (Angkiriwang et al., 2014; Hong, Lee, & Zhang, 2018);

- **Supplier constraints** - supplier constraints, also known as supply disruptions consist of situations that sometimes are unusual which can affect the supplier performance or even lead to a partial and complete failure of supply (Schmidt et al., 2012; Tinani & Kandpal, 2017). This constrains are important to be considered, so that to be mitigated (the negative effect) when they occur. Supplier constraints can be caused by factors such as earthquakes, power failures, terrorist attacks, snowstorms, customs delays, fires, slow shipments or workers strikes that can lead to shutdowns or temporary closures or causing lead-time delays due to loss of production/or transportation capability (B. He et al., 2015; Schmidt et al., 2012). Supplier backups are the most common strategies to deal with this risk (Angkiriwang et al., 2014; Hong, Lee, & Zhang, 2018);

- **Order crossover** - order crossover happens when orders are received in a different sequence from the one that they are placed (Bradley & Robinson, 2005; Chatfield & Pritchard, 2018; Hayya et al., 2009; Srivastav & Agrawal, 2018). It can occur due to two components of the replenishment lead time: the required time interval for the supplier to produce the order (which includes the actual production time, delays before production and order transmission time to the supplier) and the required time interval for the order transportation (caused by geographic location, the variability of transportation time and multiple transportation modes) (Bradley & Robinson, 2005; Chatfield & Pritchard, 2018; Srivastav & Agrawal, 2018). Several strategies are used to cope with this type of uncertainty, such as safety stock (Angkiriwang et al., 2014; Hong, Lee, & Zhang, 2018).

3.2.4 Dimensioning of safety stock (Traditional strategies)

Strategies such as safety stock and safety lead-time are typically used in inventory management to cope with both demand and supply uncertainties (C. A. Chang, 1985; Van Kampen et al., 2010). Safety stock also known as buffer stock, consists of an extra inventory held to deal with both demand and supply uncertainties so that to prevent stock-outs (Angkiriwang et al., 2014; Lambert et al., 1998; Rushton et al., 2014; N. R. Sanders, 2012; Van Kampen et al., 2010; Yamazaki et al., 2016). The safety stock of finished goods is used to attend unexpected demand, and safety stock of raw material is used to protect against supply problems and production stoppages (Lambert et al., 1998; Rushton et al., 2014; N. R. Sanders, 2012). There are multiples traditional methods for dimensioning of safety stock (See Table 3). Those methods are characterized as mathematical stochastic methods (Schmidt et al., 2012). The standard formula for calculating safety stock (Method 1) consists to multiply the safety factor (depends on the service level based on normally distributed demand) with the deviation of the demand during the replenishment time, this is, determine the safety stock as the function of service level. With the extension of this method considering the replenishment time (supplier lead time) originate the Method 2. Then, Alicke (2005) proposed a new method (Method 3) whose purpose is to determine the safety stock as the function of service level using the forecasting error for the demand during the replenishment time (determined using historical data from the mean squared deviation of the forecasted demand from the actual demand). Later, Herrmann (2011) propose Method 4 as the extension of Method 3, in which the objective was to determine the safety stock oriented to the demand through the "undershoot". Method 5 resulted from the extension of Method 4 and Herrmann (2011) extend it in order to determine Method 6, considering the "undershoot". Gudehus (2012) applied to Method 5 an adaptive service level factor, resulting in Method 7. For this, was considered that only disruption during the replenishment cycle can conduct to the absence of delivery capacity. Later on, Gudehus (2012) extends this last method to determine Method 8 by considering the dynamics of the parameters (parameters determined by means of simple exponential smoothing). The traditional Methods 1-8 (excluding Method 3) described above are based on normal distributed parameters. Lastly, Method 9 was proposed with a purpose of calculating the safety stock for a target service level of 100%, considering extreme values, mean and standard deviation (Schmidt et al., 2012).

The main methods for safety stock dimensioning described in Table 3 consider different approaches for estimating demand variability, which is a key parameter for establishing adequate safety stock levels. When assessing the applicability of the different safety stock methods in real-world supply chain contexts, we note that formulations based on the standard deviation of demand during lead time might hardly be applied (with effectiveness) in practice. This is due to the fact that demand patterns and dynamics are typically unknown and should be forecasted by a suitable forecasting approach over a given time horizon. For instance, Method 5 is widely used in seminal inventory management textbooks (P. Wang et al., 2010) and it considers stochastic demand and supply patterns. However, it does not take into consideration the variation of forecasting errors over the lead time. It is well-known that normal distribution may not be an appropriate representation of demand during the lead time because it is often skewed (Clark, 1957; Disney et al., 2016; Ruiz-Torres & Mahmoodi, 2010). Yet, we observed that several research studies have been assumed Gaussian demands in their safety stock formulations (see, for instance, Braglia et al. (2016), Caridi and Cigolini (2002a), Ruiz-Torres and Mahmoodi (2010), and Trapero et al. (2019a) and Trapero et al. (2019b)). Clark (1957) argues that the deviation of normal distribution demand during lead time can be characterized completely by the skewness. Ruiz-Torres and Mahmoodi (2010) state that "traditional models to determine the appropriate safety stock level may result in more safety stocks at sub-assembly and finished goods levels than necessary and thus lead to higher inventory carrying costs than desired. Such models generally incorrectly assume that the demand during the lead time follows a normal distribution". Disney et al. (2016) state also that, despite this is a popular approach to determine safety stock levels, it results in errors even for simple systems. An alternative is the use of Method 3, which

Table 3: Traditional methods for safety stock dimensioning (Schmidt et al., 2012; Yamazaki et al., 2016)

Method	Formula
1	$SSL = SF(SL) * \sigma_D$
2	$SSL = SF(SL) * \sigma_D * \sqrt{TRP}$
3	$SSL = SF(SL) * \sigma_F * \sqrt{TRP}$
4	$SSL = SF(SL) * \sqrt{Var(U) + TRP * \sigma_D^2}$
5	$SSL = SF(SL) * \sqrt{TRP * \sigma_D^2 + D^2 * \sigma_{TRP}^2}$
6	$SSL = SF(SL) * \sqrt{Var(U) + TRP * \sigma_D^2 + D^2 * \sigma_{TRP}^2}$
7	$SSL = SF(1 - \frac{(1-SL)*QRP}{TRP*D}) * \sqrt{TRP * \sigma_N^2 + D^2 * \sigma_{TRP}^2}, \forall QRP > TRP * D$
8	$SSL = SF(1 - \frac{(1-\alpha)*QRP}{TRP(t)*D(t)}) * \sqrt{TRP * \sigma_N(t)^2 + D(t)^2 * \sigma_{TRP}(t)^2}, \forall QRP > TRP * D$
9	$SSL = LSL_0(SL^2 - 1) + SSL_{100\%} * \sqrt{1 - (1 - SL)^c}$ $LSL_0 = \frac{QRP}{2}$ $SSL_{100\%} = \sqrt{(DV_{d,max}^+ * D)^2 + ((D_{max} - D) * TRP)^2 + (DV_{QRP,max}^-)^2}$
Legend	**SSL** - Safety Stock Level [units]; **SF** - Safety factor (depends on the service level); **SL** - Service Level; $\boldsymbol{\sigma_D}$ - Standard deviation on-demand [units/SCD]; **SCD** - Shop Calendar Day; **TRP** - Time Replenishment [SCD]; $\boldsymbol{\sigma_F}$ - the standard deviation of the forecast error for the demand during TRP [units/SCD]; **Var(U)** - Variance of the undershoot [$units^2/SCD^2$]; **D** - mean demand per period [units/SCD]; $\boldsymbol{\sigma_{TRP}}$ - the standard deviation of replenishment time [SCD]; **QRP** - replenishment quantity [units]; **TRP(t)** - replenishment time forecasted for period t [SCD]; **N(t)** - mean demand per period forecasted for period t [units/SCD]; $\boldsymbol{\sigma_N(t)}$ - the std. deviation of demand during replenishment time forecasted for period t [units/SCD]; $\boldsymbol{\sigma_{TRP}(t)}$ - the standard deviation of replenishment time forecasted for period t [SCD]; **LSL**$_0$ - lot stock level [units]; **C** - C-Norm parameter; $\boldsymbol{DV_{d,max}^+}$ - max. positive Deviation from the due date [SCD]; $\boldsymbol{DV_{max}}$ - maximum demand per period [units/SCD]; $\boldsymbol{DV_{max}}$-$\boldsymbol{DV_{QRP,max}^-}$ - max. negative Deviation in replenishment quality [units];

considers the standard deviation of forecast error during replenishment lead time (here presented as deterministic and known). However, it should be used ideally considering the Time Replenishment (TPR) as stochastic rather than deterministic, to cope with real-world supply chain needs. The main challenge inherent to their application relates to the estimation of σ_F. At this point, there are two approaches that can be followed: theoretical and empirical. The theoretical approach consists of first providing an estimation of σ_1 and then employing an analytic expression that relates σ_L and σ_1. On the other hand, the empirical approach estimates σ_L directly from the lead-time forecast error (Trapero et al., 2019a).

It is common knowledge that service levels represents a crucial input parameter for determining safety stocks. Following the described methods, the safety factor depends on the Service Level (SL). There are several ways to measure the SL, although the most discussed in the literature and therefore most common are the Cycle Service Level (CSL) - α and Fill Rate (FR) - β (Chopra & Meindl, 2016; Coleman, 2000; Jonsson & Mattsson, 2019; Ruiz-Torres & Mahmoodi, 2010; Vandeput, 2020). The CSL, also known as Type I Service Level, is defined as the probability of no stockout per replenishment cycle (i.e., portion of time between placing an order and the corresponding replenishment). The FR, also known as Volume Fill Rate (to distinguish from the Order Fill Rate) or Type II Service Level, is defined as the proportion of demand that is completely fulfilled from the available stock (Axsäter, 2015; Chopra & Meindl, 2016; Coleman, 2000; Helber et al., 2013; Vandeput, 2020). Most studies in the literature, including supply chain books, discusses the CSL measure, although, supply chain practitioners prefer the FR measure (Trapero et al., 2019a; Vandeput, 2020). Both measures have advantages and disadvantages. For instance, CSL is much easier to optimize mathematically than the FR. For computing the CSL is only necessary to consider the stock level during an order cycle, while to properly determine the FR is necessary to record the excess of demand. On the other hand, CSL does not determines the expected backorder or lost sales during a cycle. Chopra and Meindl (2016) and Vandeput (2020) argues that FR is more relevant when compared with CSL, especially when the order cycles are long. The FR is impacted by both cycle stock and safety stock, whereas the CSL is only impacted by the safety stock.

3.3 Review methodology

This review methodology represents a set of processes for selecting relevant scientific publications for this SLR. It is divided into three phases as represented in Figure 21:

- The first phase (Searching phase) involves the definition of the research query and searching for scientific publication in both Web of Science and Scopus databases;
- The second phase (Selecting phase) aims to exclude scientific publications that did not meet the defined criteria or did not address safety stock research problems;
- Lastly, the third phase (Analysing phase) consists to select relevant articles for conducting this study.

3.3.1 Searching phase

The majority of scientific publications are published in peer-reviewed scientific journals and the more relevant ones are indexed in two of the major online databases: Thomson Reuters' Web of Science (WoS) and Elsevier Scopus. The coverage of journals in WoS is approximately 13.600 journals and in Scopus is 20.346 journals (Mongeon & Paul-Hus, 2016). For this first phase of review methodology, all scientific publications are searched in both Web of Science and Scopus databases using the query described in Table 4. The search query considers keywords such as *"safety stock"* and *"safety inventory"* so that to capture in broader way topics related to safety stock problem. Keywords related to factors of uncertainty and supply chain risks in the sourcing process, such as *demand*, *price*, *lead-time*, *yield*, *order crossover*, *suppliers delay*, *variability*, *variation*, *fluctuation*, *uncertain* and *uncertainty* are also considered. Lastly, the query excludes all deterministic terms, aiming to focus only on uncertainty factors.

After performing this searching in the Scopus database resulted in a sample of 937 bibliographic references and 649 bibliographic references in the Web of Science database. All these resultant bibliographic references (from both databases) are merged and all duplicated references are removed. After that, a total of 1149 references are selected for the next phase of this review methodology.

Table 4: Query for searching of bibliographic references (Literature analysis)

Research query (Literature analysis)	((“safety stock” OR “safety inventory”) **AND** (demand **OR** price **OR** “lead time” **OR** yield **OR** “order crossover” **OR** “supplier delay” **OR** variability **OR** variation **OR** fluctuation **OR** uncertain **OR** uncertainty)) ***AND NOT** deterministic
Results in Scopus	937
Results in Web of Science	649
Results (bibliographic references merged and duplicates removed)	1149

***AND** operator is not necessary to search for bibliographic references on the WoS database

3.3.2 Selecting phase

For the selecting phase are defined three screening criteria levels in order to exclude bibliographic references that did not meet the defined criteria. For the first level of screening criteria, the choice of the consulted references was based on the following criteria:

- The bibliographic references searched included only articles from the peer-reviewed journals;
- Research articles published from 1995 to 2019, a period of 24 years;
- Publications written in English language.

In the second level of screening criteria, the SCImago Journal Rank (SJR) indicator and the subsequent journal Quartile was defined as the main selection criteria of articles for the next phase (Analysing phase). In this level of screening criteria, only articles published in journals ranked as Q1 and Q2 (Quartiles) in SJR were selected. The main objective is to consider/select relevant articles for this SLR and exclude articles that did not meet the defined criteria.

The third level of screening criteria involves the reading of the abstract of selected articles, thereby excluding articles that did not address the safety stock research problem considering at least one of risks or uncertainty factors described previously. After this phase, a total of 193 references are selected for the next phase (Analysing phase).

The co-occurrence analysis was performed in order to validate the filtering process and selection criteria of research papers (see, Section 3.4.2).

3.3.3 Analysing phase

This last phase aims to read the whole text of the article and select the more relevant ones and those that meet the purpose of this investigation. After a final manual inspection of the obtained references, a total of 193 articles was selected as the primary bibliographic reference for this SLR.

After that, all articles were classified following the safety stock research problem present in Caridi and Cigolini (2002a), therefore classified into three safety stock research problems: safety stock dimensioning, safety stock management, or safety stock positioning (allocation or placement). This classification was made by reading each article and identifying the focus of it. Some of the articles contain explicitly the research focus (research problem), but in the majority of selected articles, this classification was made exclusively through our perception where the article fits regarding the safety stock research problem.

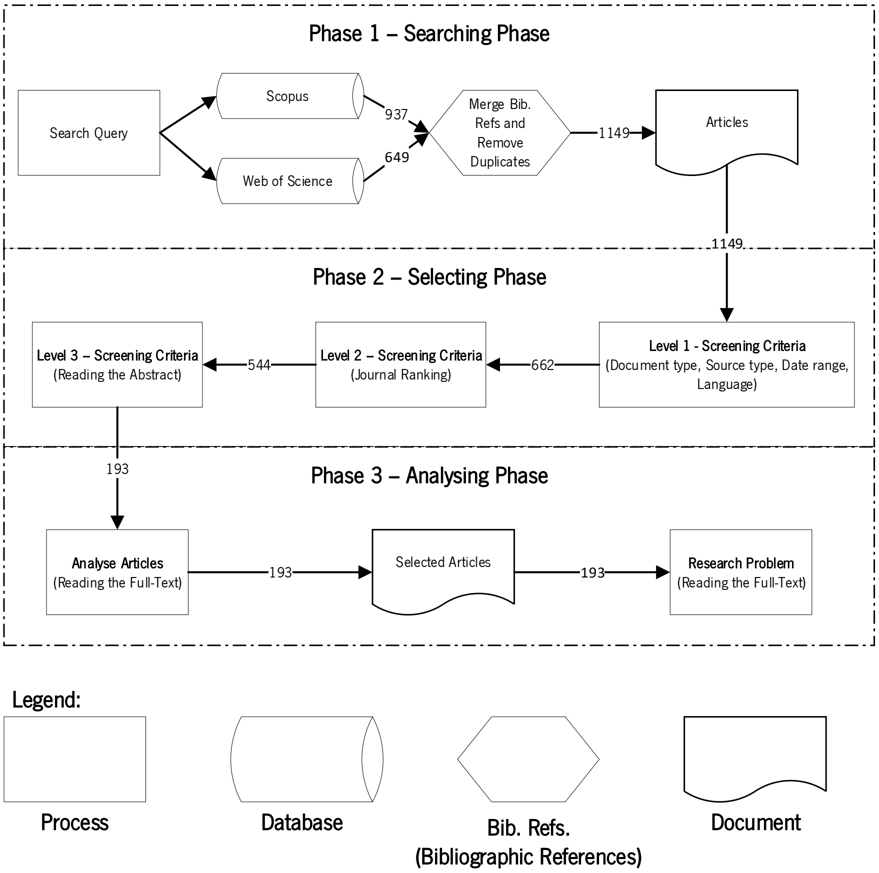

Phase 1 – Searching Phase
Search Query
Scopus
Web of Science
937
649
Merge Bib. Refs and Remove Duplicates
1149
Articles
1149
Phase 2 – Selecting Phase
Level 1 - Screening Criteria (Document type, Source type, Date range, Language)
662
Level 2 – Screening Criteria (Journal Ranking)
544
Level 3 – Screening Criteria (Reading the Abstract)
193
Phase 3 – Analysing Phase
Analyse Articles (Reading the Full-Text)
193
Selected Articles
193
Research Problem (Reading the Full-Text)
Legend:
Process
Database
Bib. Refs. (Bibliographic References)
Document

Figure 21: Adopted review methodology

3.4 Descriptive and co-occurrence analysis

3.4.1 Descriptive analysis

The descriptive analysis was performed using the **BibExcel** tool. This tool allowed to execute the initial bibliometric and statistical analysis, which included data from the Web of Science and Scopus databases (Fahimnia, Tang, et al., 2015). Then, the tool output was exported to the **Excel** tool, allowing to execute other graphical statistical analyses. The selected articles were analysed according to the number or the frequency of publications over the years, the venue of publication (name of the journal where the article is published), the research problem studied in the article, the author's influence and affiliations, and the approach adopted for modelling the problem.

3.4.1.1 Year of publication

Figure 22 illustrates the number of scientific publications published (annually) in the period from 1995 to 2019. The safety stock research problem has been gained attention from researchers especially since 2007 until now. Only 10.88% of articles were published from a period of 1995 to 1999, and 9.33% were published in the period from 2000 to 2006. From 2007 to 2019, 79.79% of articles were published, representing the increase of importance or attention of this research topic by researchers and practitioners.

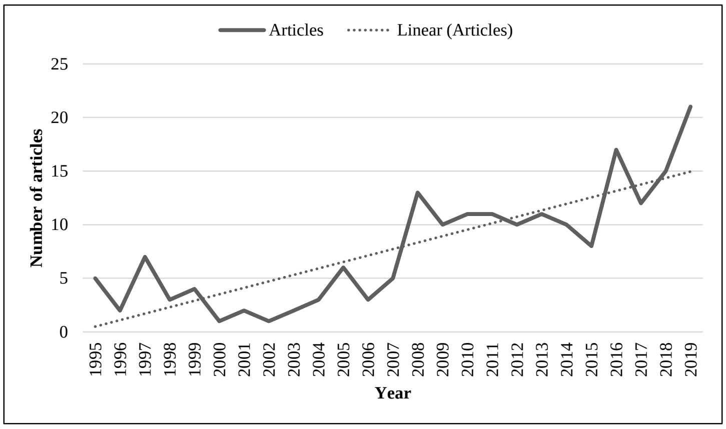


Figure 22: Distribution of scientific publication over the years

3.4.1.2 Venue of publication

Regarding the journals where the articles were published, Figure 23 shows the distribution of publications and their percentage per journal. There are 62 different journals where the reviewed articles where published. Figure 23 explicitly represents the considered journals that have at least three articles selected withing this SLR.

International Journal of Production Economics, International Journal of Production Research and European Journal of Operational Research represent the top 3 journals that mostly contributed with published articles. The first journal contributed with 41 articles that represents 21.24% of a total of reviewed articles. The second journal contributed with 14 published articles, that represents 7.25% of the reviewed articles. Finally, the third journals contributed with 13 published articles, representing 6.74% of the reviewed articles.

3.4.1.3 Research problem

The reviewed articles involve different safety stock research problem as shown in Figure 24. The problem of safety stock dimensioning is the most studied problem in the reviewed articles (a total of 79

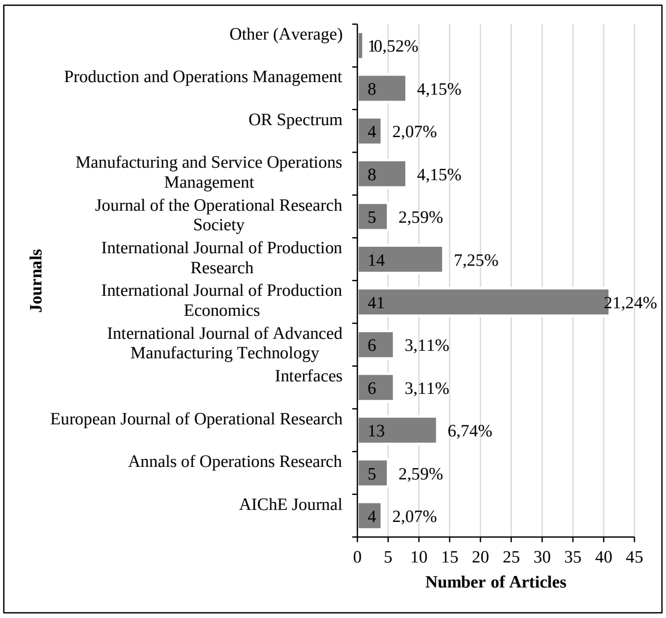


Figure 23: Distribution of publication and their percentage per journal

articles, that corresponds 40.93% of safety stock research problems covered all articles). Figure 25 and 26 illustrate the distribution of the articles for each safety stock problem in the period from 1995 to 2019.

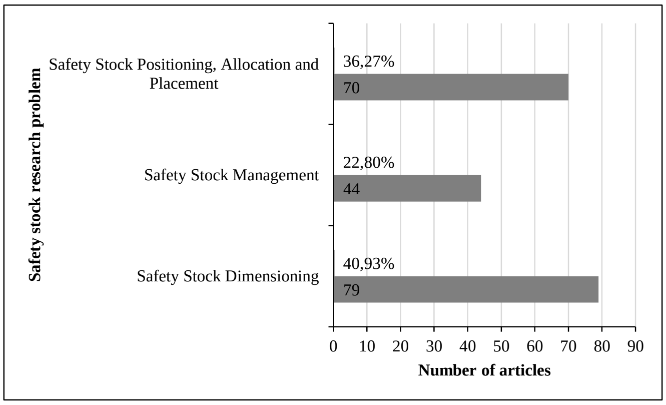


Figure 24: Distribution of publications for each safety stock research problem

3.4.1.4 Authors influence and affiliations

Table 5 describes the main authors who the most contribute with articles within the 193 articles selected. Only 26 per cent of all authors have contributed with more than one article, and the remaining 74 per cent of authors contributed with just only one research article.

The affiliation of the authors is illustrated geographically in Figure 27. Both the city and country of the author's affiliation were extracted, allowing to perform their graphical visualization using the website gpsvisualizer.com. The size of the red circle represents the occurrence of this affiliation, this is, the

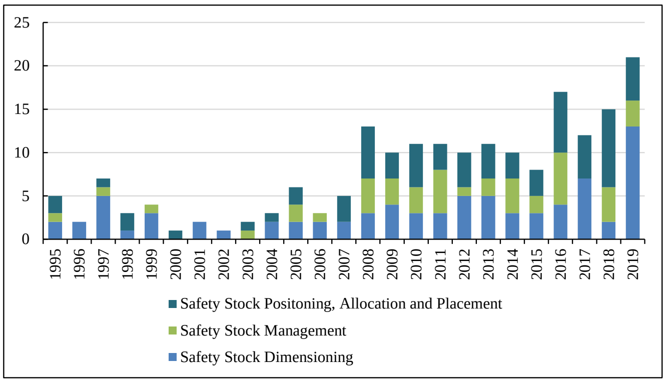


Figure 25: Distribution of publications for each safety stock research problem over the years

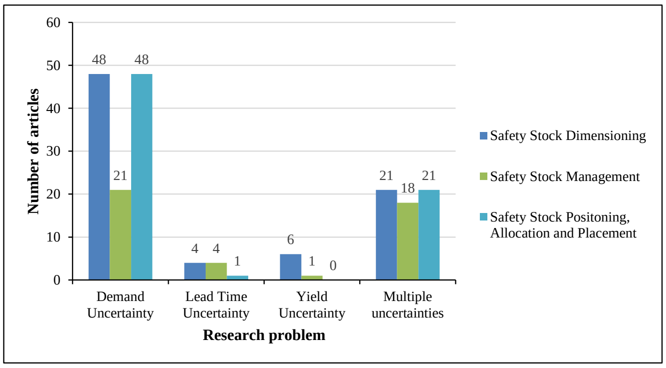


Figure 26: Distribution of publication for each safety stock problem under different uncertainty factors and risks

Table 5: Key contributing authors (first author)

Authors	Nr. of articles
Grubbström R.	4
Inderfurth K.	4
You F.	4
Braglia M.	3
Kumar K.	3
Moncayo-Martínez L.	3
Avci M.	2
Boulaksil Y.	2
Graves S.	2
Kim J.	2
Klosterhalfen S.	2
Kristianto Y.	2
Louly M.	2
Manary M.	2
Monthatipkul C.	2
Prak D.	2
Puga M.	2
Taleizadeh A.	2
Trapero J.	2
Woener S.	2

greater is the red cycle, more occurrence this affiliation have. Table 6 summarizes the number of articles published by the top contributing affiliations.

Table 6: Top contributing affiliations

Affiliation	Country	Nr. of articles
Carnegie Mellon University	United States	5
Ghent University	Belgium	5
Linköping Inst. of Technology	Sweden	5
Massachusetts Institute of Technology	United States	4
Otto-von-Guericke-Universität Magdeburg	Germany	4
Pennsylvania State University	United States	3
Purdue University	United States	3
Università di Pisa	Italy	3

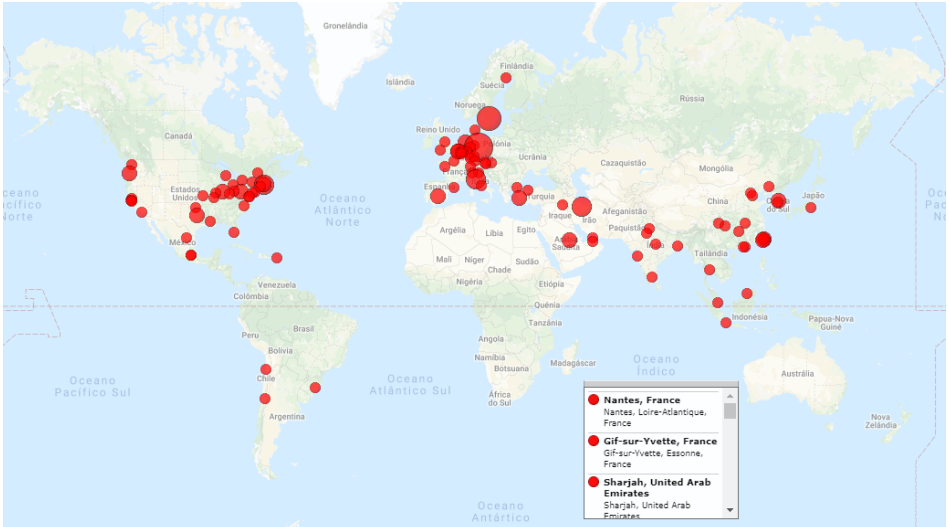


Figure 27: Geographical locations of authors affiliations (using gpsvisualizer.com)

3.4.1.5 Approach followed

In terms of the approach adopted to tackle safety stock research problems, four main approaches were used in the reviewed articles, as shown in Figure 28. Moreover, Table 7 specifies the most used techniques in the reviewed articles. In terms of the Mathematical modeling approach, the Inventory theory is the most used technique, followed by the Markov chain, Laplace transformation, Probability theory and Input-output analysis. Regarding the Optimization approach, the Heuristics technique is the most used in the reviewed articles, followed by Dynamic programming, Mixed-integer nonlinear programming, Nonlinear programming, Linear programming and Genetic algorithms (meta-heuristic). The top used Simulation techniques include Monte Carlo simulation, followed by the Discrete event simulation, Infinitesimal perturbation analysis, Event-driven simulation and Continuous simulation.

3.4.1.6 Research method

The results show that the large majority of reviewed articles (83%) used experimental research methods as the research method (see Figure 29). The case study was used in 17% of the reviewed articles. The experimental research includes methods such as simulated experiment, computational simulation or demonstration/exemplification test.

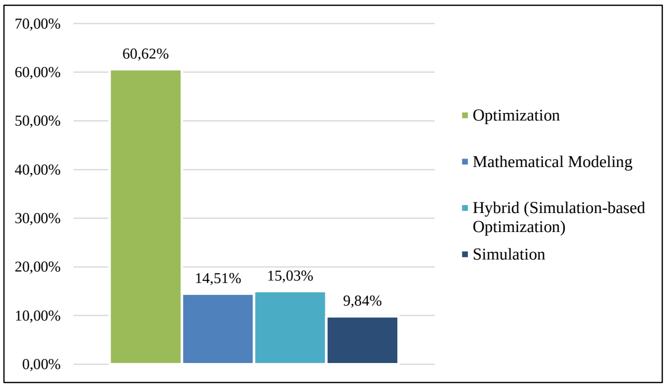


Figure 28: Distribution of adopted approaches

Table 7: Top adopted techniques

Method	Technique	Nr. of articles
Mathematical Modeling	Inventory theory	16
	Markov chain	3
	Laplace transformation	2
	Probability theory	2
	Input-output analysis	2
Optimization	Heuristics	25
	Dynamic programming	25
	Mixed-integer nonlinear programming	16
	Nonlinear programming	11
	Linear programming	9
	Genetic algorithm (meta-heuristics)	6
Simulation	Monte Carlo simulation	11
	Discrete event simulation	8
	Infinitesimal perturbation analysis	3
	Event-driven simulation	1
	Continuous simulation	1

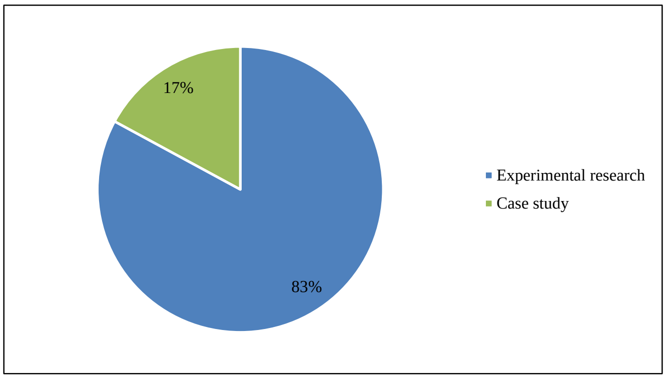


Figure 29: Distribution of research methods

3.4.2 Co-occurrence analysis

The software **VOSviewer** was used for performing this co-occurrence analysis. This tool allows the construction and visualization of bibliometric networks (van Eck & Waltman, 2010). Both of Figure 30 and 31 represent the keywords co-occurrence map of the reviewed articles. Figure 30 (left) shows the co-occurrence map of keyword after the Level 1 Screening Criteria and the Figure 31 (right) illustrate the co-occurrence map after the Level 3 Screening Criteria process. Both Screening Criteria are an integral part of phase 2 of the review methodology. The bigger circles illustrates the more occurrence of keywords in reviewed articles. The keywords with more occurrence are: *"inventory control"*, *"costs"*, *"production control"*, *"optimization"* and *"safety stock"*.

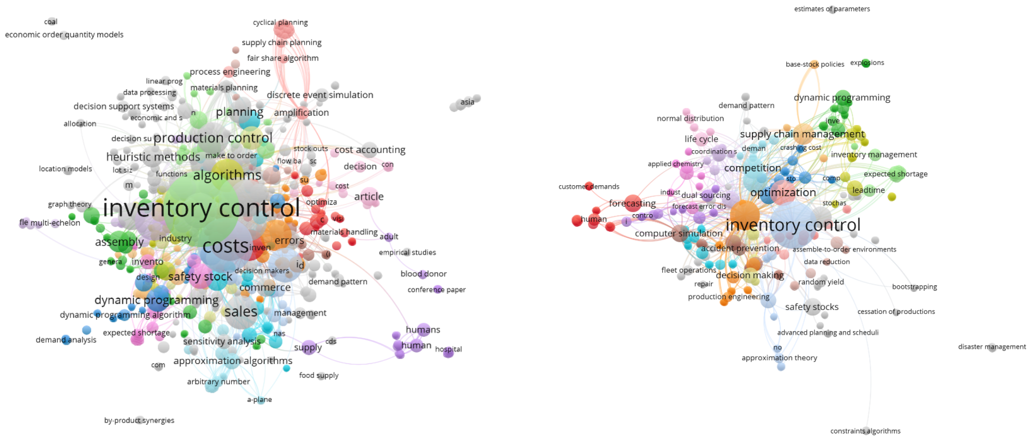


Figure 31: Co-occurrence map (Level 3)

3.5 Literature analysis (Scientific contributions)

The safety stock research problem involves typically problems of dimensioning, management and positioning, placement or allocation. Based on both safety stock research problems and the uncertainty considered in the study (multiple uncertainties or just one uncertainty factor), all the selected articles were discussed, as follows in the next sub-sections.

3.5.1 Safety stock dimensioning

Caridi and Cigolini (2002a) defined safety stock dimensioning as *"the dimensioning issue deals with finding the appropriate value of safety stocks for each item"*. In this subsection are analysed several contributions related to the safety stock dimensioning strategies under different risks and types of uncertainty.

3.5.1.1 Considering demand uncertainty

MRP is one of the most used systems for production planning and control in the manufacturing industries, helping to reduce inventory, increase operating efficiency and improve customer service. In this sense, several research studies in the literature focus on dimensioning of safety stock issue in MRP context, by considering several uncertainties/risks. Therefore, Grubbström and Molinder (1996) proposed one-level and simplest two-level serial system models to determine the optimal safety stock level using Laplace transformation and considering the traditional average cost (sum of the expected average cost of set-ups, inventory holding and backlog) as the main performance criterion. Then, three more extensions of this study were proposed. Firstly, Grubbström (1998) focused only on the one-level model, considering the Net Present Value (NPV) based criterion (the annuity streams) as the main criterion, instead of the traditional average cost approach used previously. Afterwards, Grubbström et al. (1999) generalized the models using Laplace transformations and input-output analysis, by considering demand uncertainty as Gama-distributed. Finally, Grubbström (1999) extended it for the multi-level system.

Still, in MRP environments, Zhao et al. (2001) studied and evaluated alternative methods to determine the safety stock level in multi-level MRP systems under demand uncertainty (forecast error). Others relevant studies in MRP environments can be found in Rappold and Yoho (2014). Furthermore, different studies that focus on safety stock dimensioning in Assemble-to-Order (ATO) and Make-to-Order (MTO) environments can also be found in Hsu and Wang (2001) and Jodlbauer and Reitner (2012). Hsu and Wang (2001) proposed a possibilistic linear programming model to manage production planning problems, such as the regulation of dealers forecast demand, determination of the appropriate safety stock and the number of key machines while minimizing of the sum of the product stockout costs, the material inventory holding costs, and idle capacity penalty costs. Jodlbauer and Reitner (2012) developed analytical formulas to describe the relationship between cycle time, safety stock and service level. Furthermore, they presented algorithms to find the pair cycle time and safety stock which minimize the relevant costs.

Several real-world case studies in worldwide companies have been reported (see, e.g., Caridi and Cigolini (2002b), Persona et al. (2007), Kanyalkar and Adil (2009), Boulaksil et al. (2009), Y. Chen et al. (2013), Klosterhalfen, Kallrath, and Fischer (2014), Benbitour et al. (2019) and Prawira et al. (2019)). For instance, Caridi and Cigolini (2002b) proposed and implemented a new methodology for both dimensioning and managing safety stock in an Italian leader company in the electromechanical components brand industry by considering demand forecast error as an uncertain factor. Persona et al. (2007) focused on safety stock dimensioning on both MTO and ATO environments. This study proposed models to determine optimal safety stocks for pre-assembled modules (ATO production systems) and manufacturing components (MTO production systems) used in final products. These models were applied in two Italian companies that operate in different sectors. However, Kanyalkar and Adil (2009) considered a trade-off among the plan change costs, safety stock violation penalty and inventory carrying costs for a capacitated multi-item production system in their proposed linear programming model. This model aimed at determining the optimal level of safety stock in rolling horizon. Boulaksil et al. (2009) focus on dimensioning of safety stock in multi-item multi-stage inventory system. The author proposed an approach and then implemented on a worldwide biopharmaceutical company, so-called Organon. Using the simulation based-optimization approach, Y. Chen et al. (2013) proposed a framework to determine the appropriate level of pooled safety stock levels by considering demand forecast. This framework was applied to a clinical trial company. On the other hand, Prawira et al. (2019) based on inventory control theory to proposed their model. This model focusing on determine the most reasonable amount of safety stock in the Indonesian oil and gas service companies.

Concerning the Economic Lot Scheduling Problem (ELSP) with safety stock, Brander and Forsberg

(2006) presented a model to determine the safety stock for the problem of scheduling the production of multiple items on a single facility, both with and without the existence of the idle time. Without the presence of idle time in the system, the safety stock level is calculated from the service level considering the demand variation during lead time. On the other hand, for dealing with idle time a control model is presented. In the control model, the safety stock level is calculated for time to safety stock or Time to reach the Safety Stock level (TSS).

O. Dey (2019) focused on safety stock dimensioning in single-vendor single-buyer supply chain context. This study proposed an integrated production-inventory model and also a methodology for determining the optimal values of the number of shipments from the vendor to the buyer, the safety stock, the buyer's order quantity and the probability of the production process goes *"out-of-control"*. This methodology aimed to minimize the crisp equivalent of the total cost of the integrated system. Before this research of O. Dey (2019), other studies have been conducted in this context. For instance, Glock (2012) studied a single-vendor single-buyer integrated model with stochastic demand and lot-size dependent lead time under different methods for lead time reduction (and their impact on expected total costs and safety stock). This model aimed to find the approximate optimal solution. Afterwards, Mou et al. (2017) proposed an extension of the integrated model, by considering transportation time as the main performance criteria and assuming two different safety stocks. However, is important to underline that nowadays is rarely to a supply chain operate in an environment with only one vendor and buyer.

Over times, analytical approaches have been explored to establish safety stock. For instance, Krupp (1997) proposed approaches for determining safety stock based on classic statistical theory. P. Wang et al. (2010) developed formulas to determine the reorder point and safety stock when lead time and demand are correlated. Moeeni et al. (2012) based on the basic traditional inventory models to proposed three models (for different scenarios) for determining safety stock and reorder point. Prak et al. (2017) derive closed-form expressions for the correct reorder level under uncertainty of both the mean and the variance of the demand. Moreover, both optimization and hybrid (e.g., simulation-based optimization) approaches have been also used. Hoque and Goyal (2006) developed a heuristic solution procedure to determine safety stock in an integrated inventory system under controllable lead-time between a vendor and a buyer. Srivastav and Agrawal (2016) used the Multi-Objective Particle Swarm Optimization (MOPSO) algorithm to solve their multi-objective hybrid backorder inventory model and generate Pareto curves. Huang et al. (2016) developed an optimization model to determine the optimal combination of reactive capacity and safety stock to cope with random demand, in order to minimize the total costs related to the minimum service-level constraint. Beutel and Minner (2012) developed two data-driven frameworks to determine safety stock when demand depends on external factors (e.g., prices fluctuations and weather condition). C. Zhou and Viswanathan (2011) proposed a new method for determining the safety stock under intermittent demand so-called bootstrapping method. The authors compared this new method through computational experiments with the parametric method. They concluded that the bootstrapping method works better with a large amount of randomly generated data. However, the parametric method works better with data generated in a real industry environment.

Recently, Trapero et al. (2019a) and Trapero et al. (2019b) based on empirical methods to deal with safety stock dimensioning issue. Trapero et al. (2019a) proposed empirical methods based on kernel density estimation (non-parametric) and Generalized Autoregressive Conditional Heteroscedastic (GARCH(1,1)) models (parametric) for calculating the safety stock levels under standard deviation of the lead time forecast error. On the other hand, Trapero et al. (2019b) proposed an optimal combination of the alternative empirical methods for calculating the safety stock levels, so that to minimize the piecewise linear loss function (tick loss).

Concerning of safety stock dimensioning in a production system with limited/constrained capacity, Altendorfer (2019) proposed a model for optimizing planning parameters (lot size, safety stock and planned lead time) for a multi-item single-stage production system with limited capacity. On the other hand, Helber et al. (2013), besides coping with this environment (capacity constrained production system), they also concerned with Stochastic Capacitated lot-sizing Problem (SCLSP). They proposed two different approximation models and used a fix-and-optimize algorithm to solve them, in order to determine production quantities and safety stock.

Other research studies of setting safety stock regarding Just-In-Time (JIT) production system (Ohno et al. (1995)); joint optimization of responsive supply chain design with inventory and safety stock (You and Grossmann (2008)); inventory management decision problem with service constraints (Janssens and Ramaekers (2011)); serial inventory system (Shang (2012)); periodic review inventory system with lost sales (Van Donselaar and Broekmeulen (2013)) Demand-Driven Materials Requirement Planning (DDMRP) replenishment context (Lee and Rim (2019)); supply chain reliability requirements (Lukinskiy and Lukinskiy (2017)); remanufacturing system with production smoothing (Zahraei and Teo (2018)); cyclic production schedules (Bahroun and Belgacem (2019)) have been also conducted.

This section encompasses the problem of safety stock dimensioning under demand uncertainty and comprises 48 articles (24.87% of the total sample) as described in Table 8.

3.5.1.2 Considering lead time uncertainty

Abdel-Malek et al. (2005), M. Louly and Dolgui (2009), Digiesi et al. (2013) and Sellitto (2018) are four studies that address the problem of safety stock dimensioning incorporating the lead time as the uncertainty factor (See Table 9). These four studies represent 2.07% of the total sample considered in this SLR. Abdel-Malek et al. (2005) proposed a framework based on Markovian modelling and queueing theory (tandem queues and sojourn times) that estimates the safety stock for outsourcing strategies in the multi-layered supply chain, considering lead time uncertainty. The authors highlight that in some case, long-term partnership applies better than competitive bidding/E-bidding strategies, inasmuch as the gains achieved in competitive bidding/E-bidding strategies related to the lower price and higher flexibility is dissipated by the increase of the safety stock level, and consequently the increase of inventory costs. In the context of single-level JIT assembly systems, M. Louly and Dolgui (2009) developed a novel approach based on original lower bound and dominance properties, and a branch and bound algorithm that focus only in

Table 8: Chronological scientific contributions on safety stock dimensioning under demand uncertainty

Reference	**AF**[a]	**T**[b]	**SLM**[c]	**Main criteria**
Ohno et al. (1995)	O	G	-	Min. the expected average cost per period
Adenso-Diaz (1996)	MM	ST	FR	Service level
Grubbström and Molinder (1996)	O	G	-	Traditional average cost (set-ups, holding and backlog)
Krupp (1997)	MM	IT	ND	Safety stock carrying cost; recouped profit
C. Li et al. (1997)	O	G	CSL	Min. expected annual total cost
Chan (1997)	O	H	-	-
Grubbström (1998)	O	G	-	Max. of the annuity stream
Grubbström et al. (1999)	MM	LT, IOA	-	Min. the average costs or max. of the net present value of production
Grubbström (1999)	MM	LT, IOA	-	Net present value (the annuity stream)
Hsu and Wang (2001)	O	PLP, ZFP	FR	Min. of costs
Zhao et al. (2001)	S	G	CSL	Total cost, schedule instability and SL
Caridi and Cigolini (2002b)	S	G	ND	Nr. of stock-outs, stock-out quantity and nr. of replenishments for safety buffers
Hoque and Goyal (2006)	O	H	-	Min. of the total cost, inventory holding and lead-time crashing
Brander and Forsberg (2006)	S	ND	CSL	Min. of the total costs
Persona et al. (2007)	O	ND	CSL	Min. of the total cost
Reichhart et al. (2008)	S	MCS	ND	Service level
You and Grossmann (2008)	O	MINLP	FR	Max. the net present value and min. the expected lead time
Kanyalkar and Adil (2009)	O	LP	CSL	Min. the overall cost
Boulaksil et al. (2009)	S	ND	FR	Min. total costs (holding & backorder)
P. Wang et al. (2010)	MM	IT, PT	-	-
Janssens and Ramaekers (2011)	O	LP	-	Lost sales; probability stock-out during LT
C. Zhou and Viswanathan (2011)	S	ND	CSL	Total inventory-related cost, average inventory level, fill rate and stock out rate
Feng et al. (2011)	SO	SP, IPA	FR	Min. total inventory holding and shortage costs
Jodlbauer and Reitner (2012)	O	G	FR	Min. the total relevant cost
Beutel and Minner (2012)	O	LP	CSL, FR	Min. the service level and costs
Glock (2012)	SO	G	-	Min. the expected total costs
Moeeni et al. (2012)	MM	IT	ND	Service level
Shang (2012)	O	H	-	Min. the total cost
Van Donselaar and Broekmeulen (2013)	MM	IT	FR	Fill rate
Y. Chen et al. (2013)	SO	DES, MILP	CSL	Min. the operational cost
Helber et al. (2013)	O	MILP, PLA	P	Min. the expected costs
Klosterhalfen, Kallrath, and Fischer (2014)	O	MILP	CSL	Min. the total direct rail car cost and the number of rail car types
Rappold and Yoho (2014)	O	ND	-	Min. the long-run expected costs
Gansterer et al. (2014)	SO	DES, VNS, RSM, OQ	FR	Service level
Srivastav and Agrawal (2016)	O	MOPSO, MOGA	FR	Min. the total cost, stockout units and the frequency of stockouts
Huang et al. (2016)	O	ND	CSL	Min. long-run average cost
Prak et al. (2017)	O	ND	CSL	-
Mou et al. (2017)	O	ND	-	Min. the expected cost
Lukinskiy and Lukinskiy (2017)	MM	PT	-	Total cost
Zahraei and Teo (2018)	O	NLP	-	Min. the expected total cost
Benbitour et al. (2019)	SO	DES	CSL	Min. the inventory holding and rush ordering costs
Altendorfer (2019)	O	H	ND	Min. inventory and backorder costs
Prawira et al. (2019)	O	ND	ND	Min. the costs (storage and inventory ordering costs)
Lee and Rim (2019)	MM	IT	CSL	Average inventory level and shortage rate
Trapero et al. (2019a)	S	ND	CSL	-
Trapero et al. (2019b)	O	MCS	CSL	Min. the tick loss function
O. Dey (2019)	MM	FRV	-	Min. the crisp equivalent of the expected annual integrated total cost
Bahroun and Belgacem (2019)	S	MCS	ND	Min. the safety stock and holding costs, and improving the service level

[a] **Approach followed (AF)**: MM - Mathematical modeling, O - Optimization, S - Simulation, SO - Simulation-based optimization.

[b] **Technique (T)**: DES - Discrete event simulation, FRV - Fuzzy random variable, G - Generic procedure, H - Heuristics, IOA - Input-output analysis, IPA - Infinitesimal perturbation analysis, IT - Inventory theory, LP - Linear programming, LT - Laplace transformation, MCS - Monte Carlo simulation, MILP - Mixed-integer linear programming, MINLP - Mixed-integer nonlinear programming, MOGA - Multi-objective genetic algorithm, MOPSO - Multi-objective particle swarm optimization, NLP - Nonlinear programming, OQ - OptQuest, PLA - Piecewise linear approximation, PLP - Possibility linear programming, PT - Probability theory, RSM - Response surface methodology, SP - Stochastic programming, ST - Statistical Theory, VNS - Variable neighborhood search, ZFP - Zimmermann's fuzzy programming.

[c] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed, P - Proposed service level measure.

determining the optimal safety stock of components under lead time uncertainty.

Regarding a real-world case study, Digiesi et al. (2013) proposed an extension of Sustainable Order Quantity (SOQ) model by considering lead time uncertainty and external cost of freight transport in order to identify optimal order quantity, reorder level and safety stock. A procedure was also developed to solve this model and applied to a spare parts inventory from the automotive industry.

Last but not least, Sellitto (2018) developed a method to calculate the lead-time, inventory and safety stock in a MTO job-shop manufacturing context.

Table 9: Chronological scientific contributions on safety stock dimensioning under lead time uncertainty

Reference	AF[a]	T[b]	SLM[c]	Main criteria
Abdel-Malek et al. (2005)	S	ND	ND	Annual cost
M. Louly and Dolgui (2009)	O	BB	ND	Min. the average holding cost
Digiesi et al. (2013)	O	G	ND	Logistics cost
Sellitto (2018)	SO	ND	-	-

[a] **Approach followed (AF)**: O - Optimization, SO - Simulation-based optimization.
[b] **Technique (T)**: BB - Branch and bound algorithm, G - Generic procedure, ND - Non-disclosed.
[c] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

3.5.1.3 Considering yield uncertainty

This section describes the scientific research regarding the problem of safety stock dimensioning under yield uncertainty. In this subject, there are only 6 articles (3.11% of the total sample considered) which proposed their models, approaches or frameworks following different approaches for solving this problem of safety stock dimensioning (see Table 10). In the research study of Sana and Chaudhuri (2010), a framework of production policy was developed to determine the required quantity (this is, the optimal value) of safety stock, production rate and production lot size to minimize the total expected system costs, considering machine breakdown as an uncertainty factor.

In the context of the manufacturing environment with imperfect/defective products, Taleizadeh et al. (2017) proposed an integrated inventory model for determining the optimal lot size and production uptime under random machine breakdown. The safety stock was used in the proposed model to prevent shortages in the case of machine breakdown. Recently, a similar study on manufacturing environment with defective production was conducted by Sarkar and Sarkar (2019). The study was conducted to obtain the optimal safety stock level, optimal controllable production rate and the optimal amount of production quality during the random machine breakdown under optimum energy consumption within the framework of smart production management. A real-world case study from the mining sector could be found in Song (2017). In this study, the authors proposed a new real options method (modified real options method) for determining the safety stock of ore for mining production from Kittilä mine. By comparing both this new method and the conventional Economic Order Quantity (EOQ) methods, they highlight that the real options method provides higher accuracy, better profits and robust performance when procurement costs are changed. Other relevant scientific contributions on safety stock dimensioning in the manufacturing context or contribution that consider the safety stock dimensioning as one of the multiple features for solving production/manufacturing problems, can be found in Martinelli and Valigi (2004) and de Armas and Laguna (2019).

Table 10: Chronological scientific contributions on safety stock dimensioning under yield uncertainty

Reference	AF[a]	T[b]	SLM[c]	Main criteria
Martinelli and Valigi (2004)	MM	MC	-	Min. of the average demand loss/backlog cost
Sana and Chaudhuri (2010)	MM	MC	-	Min. the total expected system cost
Song (2017)	O	RO, DP	ND	Max. the profits
Taleizadeh et al. (2017)	O	ND	-	Min. the total costs
Sarkar and Sarkar (2019)	O	ND	-	Min. the costs
de Armas and Laguna (2019)	SO	MIP, MCS	CSL	Max. the total production amount

[a] **Approach followed (AF)**: MM - Mathematical modelling, O - Optimization, SO - Simulation-based optimization.
[b] **Technique (T)**: DP - Dynamic programming, MC - Markov chain, MCS - Monte Carlo simulation, MIP - Mixed-integer programming, ND - Non-disclosed, RO - Real options technique.
[c] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

3.5.1.4 Considering multiple uncertainties and risks

Several types of research studies have been investigating the issue of safety stock dimensioning in the MRP system. In Molinder (1997), a simulation-based optimization study was proposed to jointly optimize lot-sizes, safety stock and safety lead times considering both demand and lead-time uncertainty. The author performed a comparison between safety stock and safety lead time in order to determine the best method. He highlights that both lead time and demand variability influences the level of optimal safety lead time and optimal safety stock. Furthermore, he also highlights that safety stock method is the best choice in the case of a low level of stockout/inventory holding cost ratio, high level of demand variability and low level of lead time variability. On the other hand, the safety lead time is the best choice in the case of a high level of stockout/inventory holding cost ratio and high level of demand variability. Still, Guide and Srivastava (1997) studied the dimensioning of safety stock in the MRP system modified for use in a re-manufacturing environment under random demand and lead time.

Besides MRP contexts, studies in the literature focusing on dimensioning safety stock in MPS and Available-To-Promise (ATP) environments can be found in Campbell (1995) and Hung and Chang (1999). Therefore, Campbell (1995) proposed a new method so-called "optimal safety stock"from two most known methods (constant cycle service level and constant safety stock) for establishing the safety stock in MPS environment under demand and lead time uncertainties. On the other hand, Hung and Chang (1999) considered the lead time and yield uncertainties on their proposed method for ATP environment.

Real-world case studies have been already reported by Talluri et al. (2004), Kanet et al. (2010), McNair (2015) Avci and Selim (2017), Saad, Merino Perez, and Vega Alvarado (2017) and Strohhecker and Größler (2019) regarding this issue of safety stock dimensioning by considering multiples uncertainties/risks. Thereupon, Talluri et al. (2004) applied their model for managing the made-to-stock inventories in a multinational pharmaceutical company. They considering both demand and lead time uncertainty. As a result of comparing this model with existing models, costs benefits were achieved with the proposed model. Kanet et al. (2010) proposed a software system for production planning so-called Dynamic Planned Safety Stock (DPSS) for planning a time-phased set of safety stock over a planning horizon. As a result of applying

Table 11: Chronological scientific contributions on safety stock dimensioning under multiple uncertainties and risks

Reference	UR[a]	AF[b]	T[c]	SLM[d]	Main criteria
Campbell (1995)	D, LT	MM	IT	CSL	Total cost
Guide and Srivastava (1997)	D, LT	S	ND	CSL	Stock-out percentage; SS level
Molinder (1997)	D, LT	SO	SA	ND	Total cost
Hung and Chang (1999)	Y, LT	S	LP	ND	-
Talluri et al. (2004)	D, LT	MM	IT	CSL	Min. costs and efficiency improvement
S. Chung et al. (2005)	D, LT	O	NLP	ND	Min. total cost
Katircioglu et al. (2007)	D, LT	O	G	CSL, FR	Min. expected inventory costs; max. the expected profit
Vernimmen et al. (2008)	D, LT	MM	IT	ND	Total logistics costs
Inderfurth (2009)	D, Y	MM	IT	CSL	Production, holding and shortage costs
Kanet et al. (2010)	D, LT	O	LP, G	FR	Min. total inventory and average annual fill rate
Inderfurth and Vogelgesang (2013)	D, Y	S	ND	FR	Min. the backlog and holding costs
Keskin et al. (2015)	D, Y	O	MILP, GrA, GA	-	Min. the total cost of the production plan; lot sizes
A. Kumar and Evers (2015)	D, LT	S	ND	-	Min. the total supply chain costs
McNair (2015)	D, LT	O	ARMA	ND	-
Lu et al. (2016)	D, Y	MM	IT	CSL	Service and inventory level
Chaturvedi and Martínez-De-Albéniz (2016)	D, Y	O	H	-	Min. the inventory costs
Avci and Selim (2017)	D, SD	SO	MODE/D, NSGA-II	-	Total holding cost; premium freight ratio
Saad, Merino Perez, and Vega Alvarado (2017)	D, LT	O	G	ND	Customer service, inventory and operating cost
Jonsson and Mattsson (2019)	D, LT	S	EDS	FR	Fill rate
Strohhecker and Größler (2019)	D, SC	S	SD	FR	Expected total cost
Ben-Ammar et al. (2019)	LT, OC	O	GA	-	Expected total cost

[a] **Uncertainty or risk (UR)**: D - Demand, LT - Lead time, OC - Order crossover, SC - Supplier constrains, SD - Supplier delay, Y - Yield.

[b] **Approach followed (AF)**: MM - Mathematical modelling, O - Optimization, S - Simulation, SO - Simulation-based optimization.

[c] **Technique (T)**: ARMA - Autoregressive moving average, EDS - Event-driven simulations, G - Generic procedure, GA - Genetic algorithm (meta-heuristics), GrA - Greed algorithm, H - Heuristics, IT - Inventory theory, LP - Linear programming, MODE/D - Multi-objective differential evolution algorithm, ND - Non-disclosed, NLP - Nonlinear programming, NSGA-II - Non-dominated sorting genetic algorithm II, SA - Simulated annealing, SD - System dynamics.

[d] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

this in industry, significant savings were achieved. Another real case study was reported by McNair (2015), which aimed to apply optimal safety stocks in nursing workforce management. Recently, Avci and Selim (2017) proposed multi-objective framework for supply chain inventory optimization and then developed a Decomposition-based Multi-Objective Differential Evolution algorithm (MODE/D) for this framework. This aimed to determine supplier flexibility and safety stock levels in a real-world multi-national automotive supply chain. Lastly, Saad, Merino Perez, and Vega Alvarado (2017) developed a mechanism and integrated with System Analysis Program Development (SAP) to determine adequate safety stock under the required service level. After testing the mechanism at Wavim company, the authors highlight that the mechanism should be considered as a new development for the manufacturing industry.

Inderfurth and Vogelgesang (2013) proposed an approach for determining dynamic safety stock by considering different yield uncertainties and random demand. Besides that, they presented ways to convert these dynamic safety stocks into static one, in order to be applied easily in practice. Keskin et al. (2015) proposed a Mixed-Integer Linear Programming (MILP) model to optimize simultaneously production, inventory and backorder quantities for multi-product, multi-period real-life problem by considering demand and yield uncertainties.

A. Kumar and Evers (2015) proposed an alternative approach to the random sums approach (traditional approach for determining safety stock). This alternative approach, so-called multiplication approach, consider data quality issues, as well as the correlation between demand and lead time (both stochastic) for setting safety stock.

Based on analysis from stochastic inventory control theory, Inderfurth (2009) and Lu et al. (2016) proposed their research studies for different environments. The first author, Inderfurth (2009), studied the issue of safety stock dimensioning in the production control environment, more concretely in MRP control systems by taking into consideration both demand and yield uncertainty. On the other hand, Lu et al. (2016) studied also this issue in construction material environment, considering non-stationary stochastic demand and random supply yield.

In the recent past, Chaturvedi and Martínez-De-Albéniz (2016) proposed a modelling framework based on queueing and inventory theory that optimized simultaneously inventory (safety stock), excess capacity and diversification of supply source under yield (supply capacity) and demand uncertainties. The main objective of this framework was to minimize the inventory costs (holding and shortage costs). The authors considered an infinite-horizon periodic-review inventory model for solving this problem. Recently, Ben-Ammar et al. (2019) studied the problem of multi-period supply planning. Aiming to solve this problem, a general probabilistic model under random lead-time and order crossover was proposed. Then, they developed a Genetic Algorithm (GA) for this model to determine planned lead-times and safety stock level, by minimizing expected total costs (sum of expected backlogging cost and expected inventory holding costs).

This section encompasses the problem of safety stock dimensioning under multiple uncertainties and risks and comprises 21 articles (10.88% of the total sample) as described in Table 11.

3.5.2 Safety stock management

Safety stock management is crucial for organizations so that aims to maintain customer service levels, as well as controlling the costs. Caridi and Cigolini (2002a) defined safety stock management as *"the managing issue deals with finding the appropriate time for safety stocks replenishments and with setting the appropriate delivery dates for replenishments"*. Indeed, the safety stock management intends to answer two main questions: *"when to order? And, how to order?"*. To this, there are several models in which answer in a different way to these two questions. The most known models are continuous review, periodic review and EOQ Carvalho et al. (2017).

3.5.2.1 Considering demand uncertainty

This section encompasses the problem of safety stock management under demand uncertainty and comprises 21 articles (10.88% of the total sample) as described in Table 12.

The main inventory management models are applied in environments where the demand or supply is random or uncertain. Generically, the studies proposed in the literature takes advantage of this inventory management models considering the demand as a distribution function (e.g., normal, gamma, and other) or as time series forecasting. By considering continuous-review inventory control system under stochastic demand during lead time, J. Kim and Benton (1995) studied the interrelationship between lot size and lead time and their implication on lot size and safety stock decisions (how much to order and when). The authors proposed an interactive algorithm for determining simultaneously the lot size and safety stock and then compared it with a conventional sequential approach (EOQ). As a result, they concluded that the algorithm provides better results in terms of cost savings. On the other hand, Urban (2005) developed an algorithm for solving a periodic-review problem with stochastic, serially correlated and inventory level dependent demand.

Based on optimization techniques such as MOPSO and Multi-Objective Electromagnetism-like Optimization (MOEMO), Tsou (2009) addressed the problem of multi-objective inventory control, so that to minimize the expected total cost annually under lost sales. Other optimization-based techniques, such as Mixed-Integer Non-Linear Programming (MINLP) and MILP was also used to optimize simultaneously the safety stock, reserve and base stock levels in tandem with the material flow in supply chain planning (see, Brunaud et al. (2019)).

Overall in this topic, only two studies report a real-world case study (see, You and Grossmann (2011) and Berling and Marklund (2014)). For instance, You and Grossmann (2011) developed a computational framework for simultaneously optimized the tank-sizing decisions, safety stock levels and estimated vehicle routing costs. This framework consists of stochastic approximation model (MINLP problem) under random demand.

Table 12: Chronological scientific contributions on safety stock management under demand uncertainty

Reference	AF[a]	T[b]	SLM[c]	Main criteria
J. Kim and Benton (1995)	O	G	CSL	-
Buzacott (1999)	MM	IT	ND	-
J. Kim et al. (2003)	O	G	FR	Fill rate
C. Kim et al. (2005)	SO	AVM, RL	ND	Average service level
Urban (2005)	MM	IT	-	Max. the expected profit
Lian et al. (2006)	O	NLP	ND	Min. expected cost; optimal frozen period
T. Wang and Toktay (2008)	O	H	-	Min. the expected total cost
Tsou (2009)	O	MOEMO, MOPSO	ND	Min. the expected total cost, number of stock-outs and stocked item annually
Chu and Shen (2010)	O	AA	CSL	Min. total order and holding cost
You and Grossmann (2011)	O	MINLP, BR	CSL	Min. the total expected costs
Hsueh (2011)	MM	IT	FR	Fixed manufacturing and holding costs
M. Yang and Lo (2011)	O	NLP	-	Min. the total expected inventory costs
Braglia et al. (2013)	MM	DT	CSL	Min. the expected inventory costs
Berling and Marklund (2014)	O	H	FR	Target fill rates
L. Yue et al. (2016)	O	ABC	-	Safety stock; total inventory cost
Braglia et al. (2016)	O	SA	CSL	Total cost
Torkul et al. (2016)	SO	ND	ND	Min. inventory holding cost
Turgut et al. (2018)	O	MINLP	FR	Min. the costs
Brunaud et al. (2019)	SO	MILP, MINLP	CSL	Min. transportation and inventory costs
Sakulsom and Tharmmaphornphilas (2019)	O	H	FR	Min. total inventory costs
P. Zhang et al. (2019)	MM	ND	-	Min. expected cost

[a] **Approach followed (AF)**: MM - Mathematical modelling, O - Optimization.

[b] **Technique (T)**: AA - Approximation algorithm, ABC - Artificial Bee Colony algorithm, AVM - Action-value method, BR - Branch-and-refine algorithm, DT - Diffusion theory, G - Generic procedure, H - Heuristics, IT - Inventory theory, MILP - Mixed-integer programming programming, MINLP - Mixed-integer nonlinear programming, MOEMO - Multi-objective electromagnetism-like optimization, MOPSO - Multi-objective particle swarm optimization, NLP - Nonlinear programming, ND - Non-disclosed, RL - Reinforcement learning, SA - Simulated annealing.

[c] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

From a different perspective of earlier studies in the literature regarding safety stock management problem, Hsueh (2011) considered in his research study the product life cycle (introduction, growth, maturity and decline), inventory control and manufacturing/remanufacturing system simultaneously. The author studied inventory control policies during the Product Life Cycle (PLC) and presented closed-form formulas of optimal lot size, reorder point and safety stock during each phase of the PLC. L. Yue et al. (2016) also considered the PLC and inventory control in their research. They proposed a method so-called Improved ABC-PF based on PLC theory. A PLC model based on cubic polynomial with two stages was developed and then was used Artificial Bee Colony (ABC) to optimize the parameters of the two-stage PLC model. The proposed method allows also to determine the safety stock during each PLC phase and replenishments in order to prevent stockouts.

Some of the research studies dealt with inventory control problem in multi-echelon supply chain systems. The studies proposed by Chu and Shen (2010), C. Kim et al. (2005) and Sakulsom and Tharmmaphornphilas (2019) are some examples. Chu and Shen (2010) applied a Power-of-Two (POT) policy to multi-echelon stochastic inventory model. The authors developed a polynomial-time algorithm to derive a closed-to-optimal POT policy for a given target service level. On the other hand, C. Kim et al. (2005) proposed two adaptive inventory control model (centralized and decentralized models) under non-stationary demand for solving the issue in two-echelon supply chain system (one supplier and multiple retailers). The proposed models consider the target service level predefined for each retailer as the main performance criteria. Recently, Sakulsom and Tharmmaphornphilas (2019) proposed a heuristics for determining an ordering policy in a divergent two-echelon inventory system (single warehouse and N non-identical retailers). By comparing this heuristic with Mixed-Integer Programming (MIP) models, the authors concluded that the heuristic provides goods solutions as MIP models. Based on Stock Diffusion Theory (SDT), Braglia et al. (2013) proposed a dynamic model for inventory control under non-stationary demand. The authors used the Fokker Planck (FP) equations for obtaining both the time-dependent probability distribution of the stock consumption and the reorder time.

Braglia et al. (2016) focused on safety stock management issue in a single-vendor single-buyer supply chain context under continuous review and Gaussian demand, adopting the PV criterion. They presented both approximated and exact algorithms for optimizing the safety stock.

Other studies used inventory management models considering demand as time series forecasting. Demand forecast has become an essential component in safety stock management. An inaccurate forecast can lead to inventory shortages or even overstocks and also to low customer service level. Lian et al. (2006) dealt with this issue, considering also the frozen period. They studied the frozen period in a periodic review inventory model considering forecast demands and then developed a non-linear program so-called order policy (OOP). By comparing the Forecast Order Policy (FOP) model with the proposed OOP models, they concluded that both present similar and consistent results, confirming the FOP as a very good heuristic order policy and the OOP a good alternative. Other relevant studies proposed through consecutive efforts by Buzacott (1999), T. Wang and Toktay (2008), M. Yang and Lo (2011), Torkul et al. (2016), Turgut et al. (2018) and P. Zhang et al. (2019).

3.5.2.2 Considering lead time and yield uncertainty

As shown in Table 13, there are few studied that address both the problem of safety stock management considering lead time uncertainty and considering yield uncertainty, regarding the sample considered in this SLR. A total of 4 articles (2.07%) address this problem under lead time uncertainty, namely. M. A. Louly et al. (2008), Chandra and Grabis (2008), H. Wang and Wang (2013) and Zadeh et al. (2016). Cobb (2016), is the unique study that considered the yield uncertainty in this topic (0.52% of total considered articles). For instance, M. A. Louly et al. (2008), developed a model and approach of inventory control for a single-level assembly system under random component lead times. The authors highlight that this model could be used for determining safety stock or safety lead time in the MRP context for each component under lead time uncertainty. Chandra and Grabis (2008) proposed a research study on integrating procurement costs and inventory models by considering the variable lead-time. H. Wang and Wang (2013), developed a mathematical model and deviation to determine the linkage relationship between two key parameters in inventory management: lead time uncertainty and safety stock. Focusing on addressing this problem by considering yield uncertainty, Cobb (2016) proposed an integrated inventory control model for the inspection, repair, and purchase of returnable transport items in a closed-looping supply chain. In this model, the safety stock is determined under uncertain return, so that to buffer the inventory of used and repairable containers.

Table 13: Chronological scientific contributions on safety stock management under lead time and yield uncertainties

Reference	UR[a]	AF[b]	T[d]	SLM[c]	Main criteria
M. A. Louly et al. (2008)	LT	O	BB	ND	Min. average holding cost
Chandra and Grabis (2008)	LT	O	G	-	Min. total inventory and procurement costs
H. Wang and Wang (2013)	LT	MM	MD	CSL	-
Zadeh et al. (2016)	LT	O	H	-	Min. total cost of the supply chain
Cobb (2016)	Y	MM	ND	ND	Min. expected costs

[a] **Uncertainty or risk (UR)**: LT - Lead time, Y - Yield.
[b] **Approach followed (AF)**: MM - Mathematical modelling, O - Optimization, SO - Simulation-based optimization.
[c] **Technique (T)**: BB - Branch and bound algorithm, G - Generic procedure, H - Heuristics, MD - Mathematical derivation, ND - Non-disclosed.
[d] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

3.5.2.3 Considering multiple uncertainties and risks

In the literature, many authors addressed different problems inherent to safety stock management by considering different types of uncertainties and risks. Based on optimization approaches, L. Tang et al. (2008) developed an algorithm using Lagrangian relation for solving the problem of raw material inventory faced by Shanghai Baoshan Iron and Steel Complex (Baosteel) company. The algorithm aims to determine the fixed order size and fixed interval of the replenishment process. On the other hand, for solving multi-buyer multi-vendor supply chain problem, Taleizadeh et al. (2011) proposed a harmony search algorithm to determine the reorder points, the safety stocks, and the numbers of shipments and packets in each shipment of the products under random demand and lead time. Some of these research works have dealt with dynamic inventory control policies under non-stationary demand, such as Babai et al. (2009). The authors developed an approach for the inventory control system (focused on a single-stage and single-item inventory system) under non-stationary demand incurring to the use of forecasts and random lead-time. It was developed a dynamic periodic re-order point (rk, Q) control policy for controlling the system at the end of every review period. This (rk, Q) dynamic policy was evaluated in terms of performance (service level achieved) and compared to the static (r, Q) policy. The authors concluded that both policies are similar in terms of performance.

Table 14: Chronological scientific contributions on safety stock management under multiple uncertainties and risks

Reference	**UR**[a]	**AF**[b]	**T**[c]	**SLM**[d]	**Main criteria**
Tyworth and O'Neill (1997)	D, LT	O	G	FR	Min. annual total logistics costs
L. Tang et al. (2008)	D, Y	O	LR, H	-	Min. total cost
Hayya et al. (2009)	D, LT, OC	SO	G	-	Min. costs
Babai et al. (2009)	D, LT	S	G	CSL	Total inventory costs; service level
Ruiz-Torres and Mahmoodi (2010)	D, LT	S	ND	CSL	Holding cost; Service level
Teimoury et al. (2010)	D, LT	O	MIP	ND	Min. total supply chain costs
Taleizadeh et al. (2011)	D, LT	O	HS, GA	CSL	Min. total cost
Uthayakumar and Parvathi (2011)	D, P	O	G	-	Max. profit
Ozguven and Ozbay (2012)	D, SC	MM	PVB	-	Min. total cost
Z. Zhang et al. (2014)	D, Y	O	SP, SVR	-	Min. system cost
Braglia et al. (2014)	D, LT	MM	G	ND	Min. stockholding and SS costs
S. Zhou and Chao (2014)	D, P	O	DP	-	Max. expected discounted profit
Iida (2015)	D, LT	O	DP	-	Min. expected ordering, inventory holding and shortage penalty costs.
Xiao et al. (2015)	D, P	O	DP	-	Max. expected discounted profit
Disney et al. (2016)	LT, OC	MM	IT	-	Min. inventory costs
Caceres et al. (2018)	D, LT, OC	MM	MA, MC	FR	Min. safety stock level and fill rate
Avci and Selim (2018)	D, SD	SO	MODE/D, NSGA-II	-	Min. total holding cost; inbound & outbound premium freight ratios
Chatfield and Pritchard (2018)	D, LT, OC	SO	DES, CS	CSL	Min. costs

[a] **Uncertainty or risk (UR)**: D - Demand, LT - Lead time, OC - Order crossover, P - Price, SC - Supplier constrains, SD - Supplier delay, Y - Yield.

[b] **Approach followed (AF)**: MM - Mathematical modelling, O - Optimization, S - Simulation, SO - Simulation-based optimization.

[c] **Technique (T)**: CS - Continuous simulation, DES - Discrete event simulation, DP - Dynamic programming, G - Generic procedure, GA - Genetic algorithm (meta-heuristics), H - Heuristics, HS - Harmony search, IT - Inventory theory, LR - Lagrangian relaxation, MA - Matrix analytic method, MC - Markov chain, MIP - Mixed-integer programming, MODE/D - Multi-objective differential evolution algorithm, ND - Non-disclosed, NSGA-II - Non dominated sorting genetic algorithm II, PVB - Prékopa–Vizvari–Badics algorithm, SP - Stochastic programming, SVR - Support vector regression.

[d] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

In the past few years appeared many research studies that integrated pricing and inventory models, since pricing decision has become an important issue in supply chain management. Here, S. Zhou and Chao (2014) studied this issue in a periodic-review inventory system with dual supply modes (regular and expedited) in order to mitigate the demand uncertainty under deterministic procurement costs. On the other hand, Xiao et al. (2015) focused on the effect of procurement fluctuations in the optimal pricing and sourcing policy, providing new insights related to this impact (namely to the fact that procurement cost fluctuation can alter the strategic relationship between dynamic pricing and dual sourcing, and the risk-neutral firm can achieve a higher expected profit under a more volatile spot market cost process). Other investigation that considers price sensitive (demand are auto-correlated and dependent on selling price) in their inventory model can be found in Uthayakumar and Parvathi (2011).

Z. Zhang et al. (2014) proposed an Inventory-Theory-Based Interval Stochastic Programming (IB-ISP) model for addressing the inventory problem in the electric-power generation system, considering demand uncertainty (forecast of the electricity demand) and yield uncertainty (transportation problems). The proposed model consists of planning the resources purchase patterns and electricity generation schemes of the coal-fired plants. By testing the model with real data of real environment (Beijing's electric-power generation system planning), the IB-ISP model performed better than the traditional EOQ model, since this model can provide effective measures for not-timely coal supplying pattern with reduced system-failure risk.

Recently, an real-world case study in a multi-national automotive supply chain was reported by Avci and Selim (2018) as an extension of Avci and Selim (2017). The authors developed an approach for solving the inventory replenishment problem with premium freights using simulation-based optimization techniques. They used the MODE/D for determining several parameters, such as demand forecast adjustment factor, safety stock and supplier flexibility in order to minimize the total holding cost, inbound and outbound premium freight ratios. Another real case study can be found in Ozguven and Ozbay (2012) and Teimoury et al. (2010).

Some research studies have already been proposed in the literature focusing on determining the reorder point. Hayya et al. (2009) discussed this issue considering demand uncertainty, lead time uncertainty and order crossover, where demand and lead time are independently and identically (iid) random variables. They developed regression equations for calculating the optimal cost, optimal order quantity and optimal reorder point. Another research study based on reorder point, but now considering just random demand and random lead time, was proposed by Ruiz-Torres and Mahmoodi (2010). The authors presented an alternative reorder point model (EVR method) that aimed to determine the safety stock and possible outcomes of the replenishment cycle (how much and when to replenish the inventory) without considered any distributional assumptions.

Other relevant studies proposed through consecutive efforts by Tyworth and O'Neill (1997), Braglia et al. (2014), Iida (2015), Disney et al. (2016), Caceres et al. (2018) and Chatfield and Pritchard (2018). This section encompasses the problem of safety stock management under multiple uncertainties and risks and comprises 18 articles (9.33% of the total sample) as described in Table 14.

3.5.3 Safety stock allocation, positioning or placement

In the literature, there are several terminologies for the same problem of safety stock placement. Safety stock placement, safety stock allocation and safety stock positioning represent the same problem (Graves & Willems, 2000; K. Kumar & Aouam, 2018b; H. Li & Jiang, 2012). In this SLR is adopted the terminology safety stock placement to portray this problem. The problem of safety stock placement is concerned with the question of where to position the safety stock and how much is needed (Graves & Willems, 2000). Caridi and Cigolini (2002a) defined safety stock placement as *"the positioning issue deals with finding the appropriate items in the bills of materials where safety stocks are to be placed"*.

The problem of safety stock placement is divided into two main research streams widely studied: safety stock placement for multi-stage or multi-echelon supply chain and supply chain network design with safety stock placement (Funaki, 2012). The complexity of these safety stock problems is directly related to the structure of the supply chain. There are three main structures, as depicted in Figure 32: serial network, spanning tree and general acyclic network (H. Li & Jiang, 2012; Sitompul et al., 2008). The assembly (convergent) network and distribution (divergent) network represents two special cases of spanning tree structure. The divergent network is represented with a single and central stage and several successors, and the convergent network consists of a one-end stage with several predecessors.

The serial network consists of sequential dependencies among supply chain stage, this is, each stage of the supply chain has a single predecessor and successor (H. Li & Jiang, 2012; Sitompul et al., 2008). The general acyclic network is a combination of the previous structures (Sitompul et al., 2008). There are two modelling approaches in multi-stage or multi-echelon safety stock placement: stochastic-service model and guaranteed-service model. The difference between these two approaches lies in the way that the replenishment mechanism between stages in the supply chain is modelled (Graves & Willems, 2003).

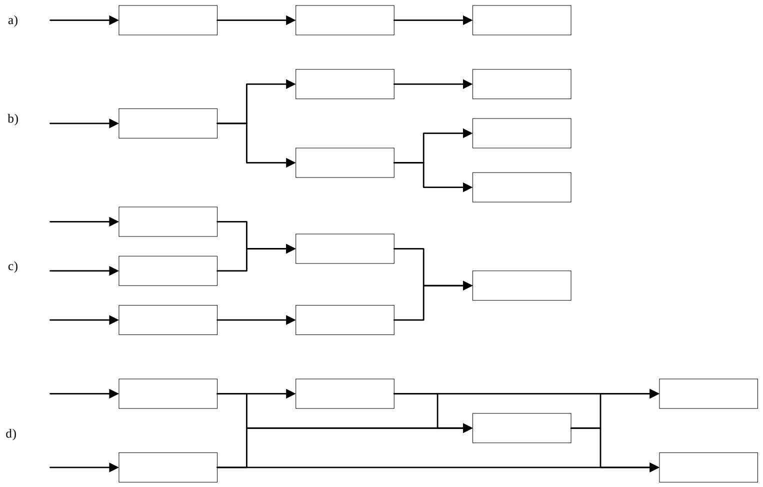


Figure 32: Supply chain structure: a) Serial network, b) Divergent network, c) Convergent network, d) General acyclic network adapted from (H. Li & Jiang, 2012)

3.5.3.1 Considering demand and lead time uncertainty

There is a set of studies in the literature regarding the problem of safety stock placement, allocation or positioning in the multi-stage or multi-echelon system and supply chain network design with safety stock placement under demand uncertainty. A total of 48 articles (24.87%) of the total sample (193 articles) address these problems. Regarding the problem of safety stock placement in the multi-echelon supply chain, several authors discussed this problem in their studies. For instance, Simpson (1958) was the first author that proposed a guaranteed-service model for supply chain structured as a serial network, so that for satisfying the demand of downstream stages at minimum inventory costs. Since then, the guaranteed service approach has been extended into several directions for solving this problem for supply chain networks modelled as assembly, distribution, spanning tree or general acyclic networks (Funaki, 2012). Inderfurth (1995) proposed a model for multi-stage supply chain structured as serial and divergent network, where demands are correlated both between products and time. The author highlight that ignoring the correlation of demand can lead to a high deviation in the optimal buffer policy. Inderfurth and Minner (1998) and Minner (1997) are an extension of Simpson's work. Both of these authors proposed a dynamic programming approaches for optimizing the safety stock in multistage inventory systems of the serial supply chain, assuming normally distributed demand and periodic review base stock control policy. Minner (1997) considered both service level and cost as performance criteria and Inderfurth and Minner (1998) assumed the service level constraints as the main performance criteria. The work proposed by Graves and Willems (2000) represents also an extension of Simpson's work. Graves and Willems (2008), Schoenmeyr and Graves (2009), Grahl et al. (2016) and K. Kumar and Aouam (2019) proposed extensions of the modelling framework developed by Graves and Willems (2000). Graves and Willems (2008) considered non-stationary demand for finding the optimal placement of safety stock under Constant Service Time (CST) policy, while Schoenmeyr and Graves (2009) considered the evolving forecast, and Funaki (2012) considered due date demand. Grahl et al. (2016) extended the approach to service time differentiation. K. Kumar and Aouam (2019) proposed a model to jointly optimize production capacity, production smoothing and service times in a multi-stage supply chain structured as spanning tree network.

Several of these research studies were applied in real-world contexts by world-wide recognized companies, such as Intel (Manary and Willems (2008) and Manary et al. (2009)), Microsoft and Case New Holland (Neale and Willems (2009)), CIFUNSA (Moncayo-Martínez and Zhang (2013)) and Teradyne, Inc. (Schoenmeyr and Graves (2009)). Other examples of real-world applications in companies operating in the automotive industry can be found in Bossert and Willems (2007), Moncayo-Martinez et al. (2014), Klosterhalfen, Minner, and Willems (2014) and Moncayo–Martínez et al. (2016) is also an example of a real-world implementation at an industrial electronics industry.

Manary and Willems (2008) developed adjustment procedures for determining the appropriate inventory target under demand uncertainty (forecast bias) for solving the problem faced by Intel in their multi-echelon inventory optimization model so-called "MEIO" regarding the presence of bias in the sales forecast data. Manary et al. (2009) extended the adjustment procedures developed in Manary and Willems (2008) considering forecast bias, non-normal forecast errors and forecast error heterogeneity.

Neale and Willems (2009) and Schoenmeyr and Graves (2009) proposed extensions of Graves and Willems (2000) for incorporating the non-stationary demand and evolving forecast. Moncayo-Martínez and Zhang (2013) proposed an extension of the Graves and Willems (2003) using meta-heuristics algorithms regarding the cost and lead time minimization of products in the generic bill of materials.

Both Moncayo-Martinez et al. (2014) and Moncayo–Martínez et al. (2016) proposed studies for addressing the problem of safety stock placement for the automotive industry. The first study used meta-heuristics or modern optimization algorithms (swarm intelligent algorithms: ant colony and intelligent water drop) and in the second study developed a framework.

Other relevant studies proposed through consecutive efforts by Sitompul et al. (2008), Albrecht (2014), H. Chen and Li (2015), Hua and Willems (2016) are found in the literature. All these studies aim to optimize the multi-echelon inventory system under guaranteed service approach.

Table 15: Chronological scientific contributions on safety stock allocation, positioning and placement under demand and lead time uncertainties

Reference	**UR**[a]	**AF**[b]	**T**[c]	**SLM**[d]	**Main criteria**
Inderfurth (1995)	D	O	G	CSL	Min. expected holding costs
Schneider and Rinks (1995)	LT	SO	G	CSL	Min. overall costs
Minner (1997)	D	O	DP	CSL	Min. average holding costs
Inderfurth and Minner (1998)	D	O	NLP	CSL	Min. of costs, service level
Shen et al. (2003)	D	O	NLP	CSL	Min. the total cost
Cao and Silver (2005)	D	O	H	-	Min. the total expected units short
Monthatipkul and Yenradee (2007)	D	O	LP	FR	Min. of the total cost
Bossert and Willems (2007)	D	O	DP	CSL	Min. inventory levels
Manary and Willems (2008)	D	O	G	CSL	Units of product
Ozsen et al. (2008)	D	O	LR	CSL	Min. the sum of facility location, transportation, inventory costs
Sitompul et al. (2008)	D	S	MCS	ND	Service level
Graves and Willems (2008)	D	O	DP	-	Min. safety stock holding costs
Kaminsky and Kaya (2008)	D	O	H	ND	Min. of costs
Manary et al. (2009)	D	O	G	ND	Min. production costs, lost-sales costs and deviation cost
Neale and Willems (2009)	D	O	ND	CSL	Min. safety stock holding cost
Schoenmeyr and Graves (2009)	D	O	DP	ND	Min. inventory holding costs
You and Grossmann (2010)	D	O	MINLP	CSL	Min. total supply chain design cost
Nasiri et al. (2010)	D	O	NLP	CSL	Total cost
Yao et al. (2010)	D	O	MINLP	CSL	Min. expected total cost
Monthatipkul et al. (2010)	D	SO	H	-	Min. lost sales
You and Grossmann (2011)	D	O	MINLP	CSL	Min. the annualized cost and the maximum GS[*]times of the markets
Tian et al. (2011)	D	O	LP	ND	Max. profits and min. safety stock costs
S. Liao et al. (2011)	D	O	MINLP, NSGAII	FR	Min. total cost, Max. fill rates, responsive level
Funaki (2012)	D	O	DP	CSL	Min. total costs
D. Yue and You (2013)	D	O	MINLP	CSL	Min. the total cost over
Moncayo-Martínez and Zhang (2013)	D	O	ACO, IWD	ND	Min. total supply chain cost and product lead time
Moncayo-Martinez et al. (2014)	D	O	DP	-	Min. safety stock cost
Albrecht (2014)	D	O	H	ND	Min. long-run average expected inventory and backorder costs
Klosterhalfen, Minner, and Willems (2014)	D	O	DP	CSL	Min. total safety stock holding cost
Petridis (2015)	D	O	MINLP	CSL	Min. cost of the supply chain
H. Chen and Li (2015)	D	O	DP	CSL	Minimization of the average total cost
Tempelmeier and Bantel (2015)	D	SO	G	FR	Min. transportation and holding costs
Grahl et al. (2016)	D	O	GA	CSL	Min. total holding costs
Moncayo–Martínez et al. (2016)	D	O	ACO, IWD	-	Min. inventory cost and lead time

[a] **Uncertainty or risk**: D - Demand, LT - Lead time.

[b] **Approach followed**: O - Optimization, S - Simulation, SO - Simulation-based optimization.

[c] **Technique**: ACO - Ant colony optimization (meta-heuristics), CQMIP - Conic quadratic mixed-integer programming, DP - Dynamic programming, G - Generic procedure, GA - Genetic algorithm (meta-heuristics), H - Heuristics, IWD - Intelligent water drop algorithm (meta-heuristics), LP - Linear programming, LR - Lagrangian relaxation, MCS - Monte Carlo simulation, ND - Non-disclosed, MINLP - Mixed-integer nonlinear programming, NLP - Nonlinear programming, NSGA-II - Non-dominated sorting genetic algorithm II, PSO - Particle swarm optimization algorithm (meta-heuristics), STA - Spanning tree-based algorithm.

[d] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

[*] GS - Guaranteed service.

The problem of supply chain network design with safety stock placement represents a classical problem in operational research (Funaki, 2012). This consists to jointly optimize design decisions of the supply chain with safety stock placement. Studies proposed by Yao et al. (2010), You and Grossmann (2010), S. Liao et al. (2011), Funaki (2012), Moncayo-Martínez and Zhang (2013) and S. Li et al. (2017) address this kind of problem. Yao et al. (2010) proposed a mixed-integer nonlinear programming model to address the facility location-allocation and inventory problem. A solution procedure was developed also to solve the proposed model. You and Grossmann (2010) presented a mixed-integer nonlinear programming model for determining the optimal transportation, inventory level and network structure in a multi-echelon supply

Table 16: Chronological scientific contributions on safety stock allocation, positioning and placement under demand and lead time uncertainties

Reference	**UR**[a]	**AF**[b]	**T**[c]	**SLM**[d]	**Main criteria**
Grace Hua and Willems (2016)	D	O	ND	-	Min. total safety stock cost
Boulaksil (2016)	D	S	G		Min. holding and backorder cost
Ross et al. (2017)	D	O	MINLP, H	CSL	Min. the total annual cost
Schuster Puga and Tancrez (2017)	D	O	CQMIP, H	CSL	Min. the location, transportation and inventory costs
S. Li et al. (2017)	D	O	MINLP	CSL	Min. the total cost
van der Rhee et al. (2017)	D	O	H	CSL	Min. the total holding cost
Shahabi et al. (2018)	D	O	MINLP	CSL	Min. the facility location, transportation and inventory cost
Hong, Dai, et al. (2018)	D	O	PSO, STA	ND	Min. the overall cost of the SC
Negahban and Dehghanimohammadabadi (2018)	D	O	MINLP	ND	Max. total net profit
Woerner, Laumanns, and Wagner (2018)	D	SO	G	FR	Min. overall holding costs
K. Kumar and Aouam (2018b)	D	SO	DP	ND	Min. system-wide production and inventory costs subject
K. Kumar and Aouam (2018a)	D	O	DP	ND	Min. WIP and holding costs; setup time reduction investment
Fichtinger et al. (2019)	D	O	ND	-	Min. total costs
Tookanlou and Wong (2019)	D	O	DP	-	Min. expected total cost
M. Kumar et al. (2019)	D	O	DP	CSL	Min. expected total cost

[a] **Uncertainty or risk**: D - Demand, LT - Lead time.

[b] **Approach followed**: O - Optimization, S - Simulation, SO - Simulation-based optimization.

[c] **Technique**: ACO - Ant colony optimization (meta-heuristics), CQMIP - Conic quadratic mixed-integer programming, DP - Dynamic programming, G - Generic procedure, GA - Genetic algorithm (meta-heuristics), H - Heuristics, IWD - Intelligent water drop algorithm (meta-heuristics), LP - Linear programming, LR - Lagrangian relaxation, MCS - Monte Carlo simulation, ND - Non-disclosed, MINLP - Mixed-integer nonlinear programming, NLP - Nonlinear programming, NSGA-II - Non-dominated sorting genetic algorithm II, PSO - Particle swarm optimization algorithm (meta-heuristics), STA - Spanning tree-based algorithm.

[d] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

[*] GS - Guaranteed service.

chain. S. Liao et al. (2011) in its turn, proposed a mixed-integer programming model for multi-objective optimization of the supply chain network and a multi-objective evolutionary algorithm approach. This model considers the total cost, customer service level (fill rate) and flexibility as the main performance criteria.

Funaki (2012) proposed a multi-echelon safety stock placement model in supply chain design under due-date demand and an optimization procedure for this model. This multi-echelon safety stock placement model represents an extension of guaranteed-service model proposed in Graves and Willems (2000) and Inderfurth and Minner (1998). Moncayo-Martínez and Zhang (2013) developed an approach based on the MAX-MIN ant system for solving safety stock placement problem in which to minimize the total supply chain cost and product lead time. Other relevant studies regarding the problem of supply chain network design with safety stock placement can be found in Petridis (2015), K. Kumar and Aouam (2018b) and Fichtinger et al. (2019).

Lastly, Cao and Silver (2005), Kaminsky and Kaya (2008), Ozsen et al. (2008), Nasiri et al. (2010), You and Grossmann (2011), Tian et al. (2011), D. Yue and You (2013), Tempelmeier and Bantel (2015), Boulaksil (2016), Ross et al. (2017), Schuster Puga and Tancrez (2017), van der Rhee et al. (2017), Shahabi et al. (2018), Negahban and Dehghanimohammadabadi (2018), Hong, Dai, et al. (2018), Woerner, Laumanns, and Wagner (2018) and Tookanlou and Wong (2019) also proposed their model for solving safety stock placement problem considering demand as uncertainty factor.

Only 0.52% (1 article) address this problem of safety stock placement considering uncertain lead-time. Schneider and Rinks (1995) proposed approximations of two-echelon periodic review inventory model under lead time uncertainty, using the power approximation. Table 15 and 16 provides an overview of all articles considered to this topic of safety stock placement under demand and lead time uncertainty. This overview includes the description of the type of uncertainty considered by the author, as well as the approach followed and the main performance criteria.

3.5.3.2 Considering multiple uncertainties and risks

Regarding the problem of safety stock placement for multi-stage or multi-echelon supply chain, Simchi-Levi and Zhao (2005) proposed a framework for evaluating and coordinating inventory policies for supply chains with three network structures (serial, assembly and distribution systems) following the stochastic service model approach and considering demand and lead time uncertainties. Each stage of this structure controls its inventory with continuous base-stock policy. Osman and Demirli (2012) proposed a safety stock placement models (decentralized and centralized) for determining and placing the safety amounts in a multistage supply chain under demand and lead time uncertainty. The fill rate and safety stocks at each stage of the supply chain are determined in order to minimize the safety stock placement costs through the entire supply chain. Unlike studies proposed in Humair and Willems (2011) and Willems (2008) where the guaranteed service model for general acyclic supply chain was extended considering the demand uncertainty and deterministic lead time, Humair et al. (2013) extended the guaranteed service model incorporating both demand and lead time uncertainties. The guaranteed service model proposed in Hua and Willems (2016) applied the study of Graves and Willems (2005), but for two-stage serial line supply chain instead for the spanning tree network as considered by these authors. This configuration model aimed for determining the chosen option (cost and lead time pairing) and inventory stocking level at each stage of the supply chain under demand and lead time uncertainty. Graves and Schoenmeyr (2016) generalized the guaranteed-service model for safety stock placement incorporating the yield (capacity constraints) and demand uncertainties. Other relevant studies regarding the problem of safety stock placement under multiple uncertainties/risks can be found in Sonntag and Kiesmüller (2017), Woerner, Laumanns, and Wagner (2018) and De Smet et al. (2019). Sonntag and Kiesmüller (2017) proposed a model of in-house multi-stage serial production systems with random yield and demand, in order to calculate the optimal safety stock and positions of quality inspections through the production stages and optimize the position of inspections. Woerner, Laumanns, and Wagner (2018) developed a simulation-based optimization model for determining the optimal base stock level of a multi-echelon assembly system under capacity constraints (yield uncertainty) and uncertain demand. The authors compared this model with guaranteed service model providing better results in terms of reducing costs keeping the same service level. Last but not least, De Smet et al. (2019) proposed two modelling approaches (extensions of the guaranteed-service models and the stochastic service model) for multi-echelon inventory optimization problem in a distribution network under lead time and demand uncertainties.

Recently, Schuster Puga et al. (2019) addressed a study related to the problem of supply chain network design with safety stock placement. They formulated a model for two-stage supply chain design for jointly integrate the safety stock placement and delivery strategy decisions considering demand and lead time uncertainties, in order to minimize the costs of transportation, facility opening, cycle inventory, ordering and safety stocks. In this work, the guaranteed-service approach was used for modelling the safety stock placement decisions.

Other relevant scientific studies regarding this topic of safety stock placement can be found in Swaminathan and Tayur (1998), Lin et al. (2000), Shen et al. (2003), Bollapragada et al. (2004), Vanteddu et al. (2007), Jung et al. (2008), Desmet et al. (2010), Epstein et al. (2012), Kristianto et al. (2012), Kristianto and Zhu (2013), X. Xu et al. (2016) and D. Kumar and Kumar (2018). This topic related to the problems of safety stock placement has been widely studied by several authors, considering different types of uncertainty factors, performance criteria and following different modelling approaches, as shown in Table 17. Considering the total sample of considered articles (193) to this SLR, 21 of these articles (10.88%) addressed this problem.

Table 17: Chronological scientific contributions on safety stock allocation, positioning and placement under multiple uncertainties and risks

Reference	UR[a]	AF[b]	T[c]	SLM[d]	Main criteria
Swaminathan and Tayur (1998)	D, Y	O	SP	-	Min. sum of stockout and holding costs
Lin et al. (2000)	D, LT, SD	O	NLP	CSL	Min. total inventory capital
Bollapragada et al. (2004)	D, LT, Y	SO	H, MCS	CSL	Min. costs
Simchi-Levi and Zhao (2005)	D, LT	SO	DP, MCS	FR	Min. inventory cost
Vanteddu et al. (2007)	D, LT	O	ND	ND	Total safety stock cost
Jung et al. (2008)	D, Y	SO	LP, DES	ND	Min. total expected inventory
Desmet et al. (2010)	D, LT	SO	DES	FR	Min. components fill rate
Epstein et al. (2012)	D, Y	O	G	ND	Min. global empty container costs
Osman and Demirli (2012)	D, LT	O	BD	FR	Min. safety stock costs
Kristianto et al. (2012)	D, LT	SO	GA, MCS	ND	Backorders and inventory level
Kristianto and Zhu (2013)	D, LT	O	H	ND	Nr. of backorders, safety stock level and lead time variability
Humair et al. (2013)	D, LT	O	NLP	ND	Min. inventory costs
X. Xu et al. (2016)	D, LT, SD	O	DP	-	Min. total safety stock and project cost
Hua and Willems (2016)	D, LT	O	ND	-	Min. total supply chain costs
Graves and Schoenmeyr (2016)	D, Y	O	DP, H	-	Min. holding cost
Sonntag and Kiesmüller (2017)	D, Y	O	H	CSL	Min. overall costs
Woerner, Laumanns, and Wagner (2018)	D, Y	SO	CLM, IPA	FR	Min. holding costs
D. Kumar and Kumar (2018)	D, LT	S	SysD	CSL	Min. inventory costs
Ghafour (2018)	D, LT	O	G	ND	Min. total cost
Schuster Puga et al. (2019)	D, LT	O	CQMIP	CSL	Min. overall costs
De Smet et al. (2019)	D, LT	O	G	FR, CSL	Min. holding, fixed order and operating flexibility costs

[a] **Uncertainty or risk (UR)**: D - Demand, LT - Lead time, SD - Supplier delay, Y - Yield.

[b] **Approach followed (AF)**: O - Optimization, S - Simulation, SO - Simulation-based optimization.

[c] **Technique**: BD - Benders decomposition, CLM - Constrained level method, CQMIP - Conic quadratic mixed-integer programming, DES - Discrete event simulation, DP - Dynamic programming, G - Generic procedure, GA - Genetic algorithm (meta-heuristics), H - Heuristics, IPA - Infinitesimal perturbation analysis, LP - Linear programming, MCS - Monte Carlo simulation, MINLP - Mixed-integer nonlinear programming, ND - Non-disclosed, NLP - Nonlinear Programming, NLP - Nonlinear programming, SP - Stochastic programming, SysD - System Dynamics.

[d] **Service level measure (SLM)**: CSL - Cycle service level, FR - Fill rate, ND - Non-disclosed.

3.6 Literature gaps and research opportunities

Considering the literature analysis described in the Section 3.5, we identify in this section some research gaps. Moreover, research opportunities and a Literature Map are also provided.

Figure 34 illustrates the Literature Map resulted from this SLR, providing an overview of distinct proposed formulations of the safety stock problem in the literature. The Literature Map consists of four levels of iterations. The first level shows the literature gaps identified in the literature. The second level describes the safety stock problems, namely Safety stock dimensioning, Safety stock management and Safety stock placement (allocation or positioning). The third level describes several uncertainties factors and risks associated with the procurement process and therefore considered as input to address safety stock related problems, namely Demand uncertainty, Lead-time Uncertainty, Yield uncertainty and Multiple uncertainties and risks. The last level represents different approaches followed, as well as scientific contributions that use these same approaches to solve safety stock related problems.

Analysing Figure 34 we highlight that:

- In general, the Optimization approach is the most used to address safety stock problem and techniques such as heuristics, dynamic programming and mixed-integer nonlinear programming are the most used techniques related to the Optimization approach (as described in Table 7).
- Demand uncertainty is the most common uncertainty factor in the proposed inventory models. On the other hand, there is a lack of studies that considered the lead time uncertainty, as well as the yield uncertainty.
- Recent data-driven approaches, such as BA and Big Data Analytics (BDA), are producing a strong impact in diverse research fields, including supply chain management. However, BA and BDA has not yet been explored to solve safety stock related problems.

After conducting a critical literature analyses, described in section 3.5, some research gaps are identified and discussed herein, as well as the research opportunities:

- Several studies in the literature and leading supply chain books, as well as inventory management software tools, assume that the demand during the lead time follows a normal distribution. Yet, several authors have already warned that such assumption may be flawed because lead time demand is often skewed (see, Janssens and Ramaekers (2011), Lee and Rim (2019), and Ruiz-Torres and Mahmoodi (2010)). This statistical assumption can lead to higher service level than desired, resulting an overestimation of safety stock and consequently higher inventory costs (Ruiz-Torres & Mahmoodi, 2010). Hence, in practice, future demands must be forecasted based on historical observations.
- The majority of peer-reviewed articles focus on determining safety stock/inventory based on statistical parameters (e.g., standard deviation or mean of demand) and simplifications (e.g., distribution

of statistical parameters, parameters are known) (Schmidt et al., 2012). There is a lack of articles that focus on providing dynamic models that consider the knowledge of future volatility of parameters for determining safety stock. More research is needed to explore not only the application of more realistic safety stock closed-form stochastic approaches considering the variation of forecasting errors rather than the variation of demand, especially in multi-product multi-echelon inventory management settings, but also to study the benefits of such safety stock methods in case studies with practical interest. Moreover, empirical non-parametric approaches for estimating the variability of forecast errors (see, Trapero et al. (2019a), Gonçalves et al. (2021) and Trapero et al. (2019b)) could be further exploited using for example business analytics techniques. Note that BA and BDA techniques allow the use of predictive analytics for applying machine learning techniques on real data in order to learn or obtain knowledge from data and predict future supply chain demand based on historical and current data. In this context, the prediction capabilities could be also optimized using metaheuristics.

- Several methods for calculating safety stock can be found in the literature based on two main service level measures, namely cycle service level and fill rate. Although the cycle service level has been criticized for not being relevant from a customer perspective and also not recommended for inventory control practice (Axsäter, 2015; Jonsson & Mattsson, 2019), it remains the most used in the literature, as illustrated in Figure 33. Several studies and supply chain books considered the CSL measure because, unlike the FR measure, it is of easy computation. Chopra and Meindl (2016), Tyworth (1992), and Vandeput (2020) argued the necessity of transition from CSL to FR because fill rate is a more relevant measure.

- The inventory control problem under lead-time uncertainty is not sufficiently studied, particularly in assembly networks (M. A. Louly et al., 2008). Demand uncertainty is the most considered factor in the literature (see Figure 26 and 34). Several studies considered constant lead time, which is not realistic for major of supply chain environments due to the unexpected events that can occur causing random delays. These delays may require to incurring the special/premium freight so that to avoid stockouts and consequently an extra cost for organizations. Moreover, there are few studies that address the safety stock problems on MRP environment considering the lead time uncertainty. Herein, empirical non-parametric approaches could also be exploited to address lead-time uncertainty.

- The impact of order crossover in determining safety stock is under-researched. Recent studies in the literature demonstrated not considering order crossover can be translated to larger inventory costs (Chatfield & Pritchard, 2018). Riezebos (2006) argued that the modern supply chain needs to address the issues concerning expected order crossovers, generally neglected in inventory control literature. Modern supply chains facing the growing occurrence of order crossover, as well as with the increasing importance of service performance (Chatfield & Pritchard, 2018). Chatfield and

Pritchard (2018) stated that "classical inventory modelling methods should be re-examined and perhaps reformulated in order to accommodate the possibility of order crossover".

- There is a lack of research studies that address the safety stock problem by considering the variation of demand over the PLC and seasonality (Strohhecker & Größler, 2019). In this review, the only works addressing this issue are Hsueh (2011) and L. Yue et al. (2016). The PLC is becoming smaller due to technological advances as happens for instance in the mobile phone and electronics components industries. During the PLC, the product demand may increase rapidly at the ramp-up stage, then it stabilizes and starts decreasing at the decline stage. Several traditional inventory models considered that this increase of product demand as stationary, instead of a change in demand in a certain stage of the product life cycle. An accurate demand forecasting is crucial for real-world application (directly effects the safety stock level, as well as the total inventory costs) and sometimes is very difficult to be estimated under short product life cycle (for instance, fashion products such as shoes and clothing). Techniques such as BA and BDA could be helpful to cope with this issue. For instance, Huang et al. (2016) highlight that companies can take advantage of big data for coping with demand surges. In effect, BDA is producing a great impact in various research fields including SCM, providing tools for supporting and enabling strategic and operational decision-making.

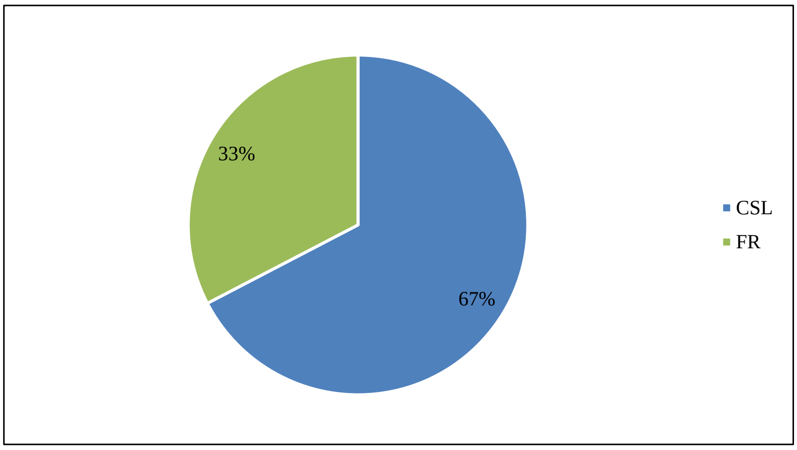


Figure 33: Distribution of Service level measure adopted

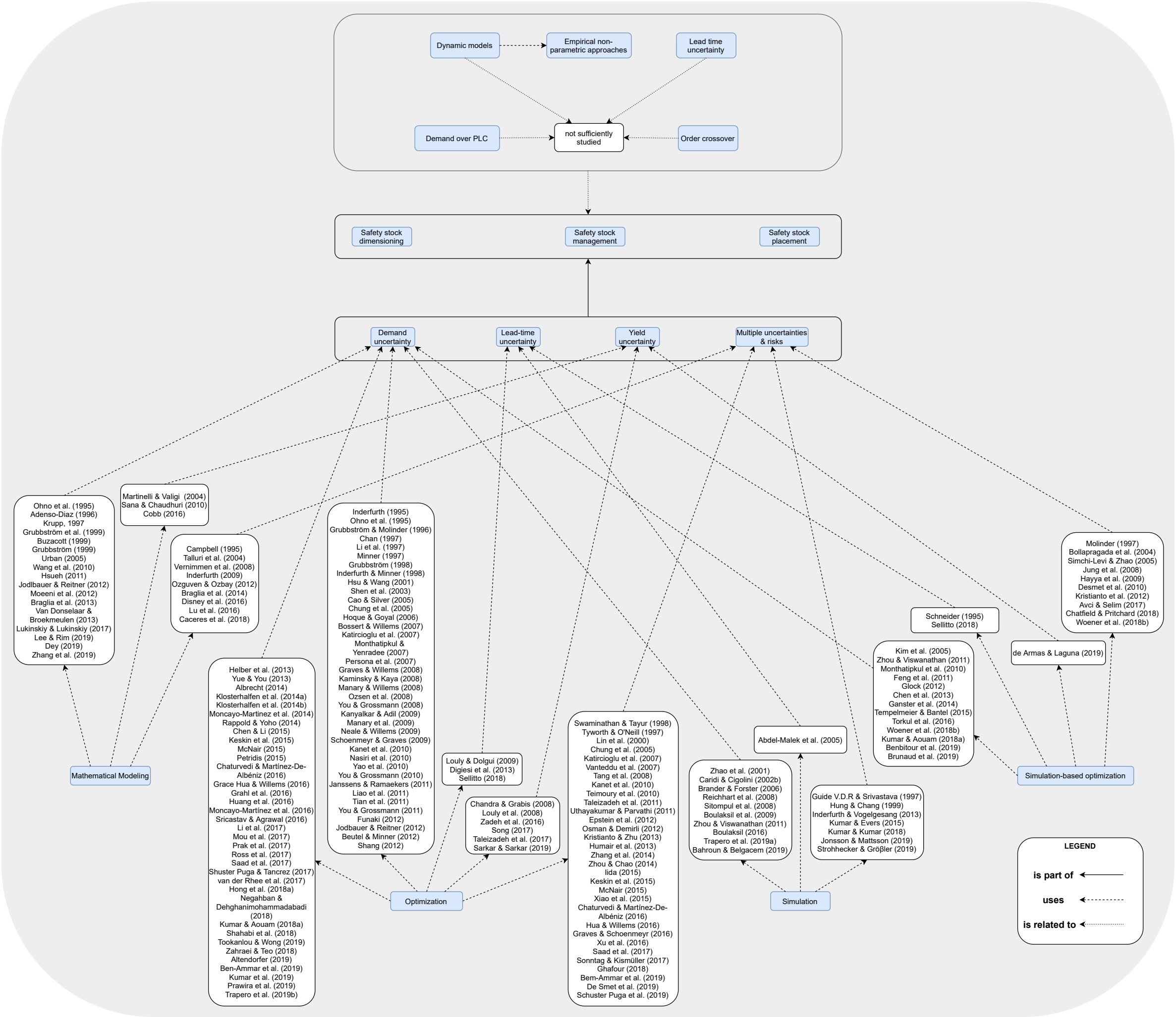


Figure 34: Literature Map

3.7 Summary and directions for future research

In this paper, we review the topic of safety stock dimensioning strategies under uncertainty factors in procurement process. Safety stocks are important at all stages of the supply chain and due to this makes it an attractive field for researches and practitioners. This topic has been gaining increasing attention over time and this trend is confirmed with the increasing number of publications (see Figure 22). The systematic literature review was performed following a review methodology which represents a set of processes for selecting relevant scientific publications. It starts with the definition of the *"search query"* that is applied in both WoS and Scopus databases (major online databases where the relevant peer-reviewed scientific journals are indexed). After this first stage, the scientific publications are filtered and finally grouped into three main safety stock research domains: safety stock dimensioning, safety stock management and safety stock allocation or positioning or placement. As a result, a set of 193 scientific publications was selected from 1995 to 2019. A co-occurrence analysis is performed in order to identify research concepts related to the safety stock problem. This review might have limitations, even with a large number of scientific publications analysed (is not devoid of limitations). Two main limitations are pointed out. Firstly, some of the relevant publication could be non-identified due to the *"search query"* developed in the review methodology. Secondly, we only considered publications that meet the defined criteria in three *"screening criteria"* of our review methodology (e.g., only considered peer-reviewed publications while excluding the conference proceedings and only considers publication written in English). The current research gaps and research opportunities are identified and discussed so that to provides a road map to guide future research agenda on this topic.

Considering the presented literature review, we highlight several relevant insights regarding different contexts:

- For the MRP context under both demand and lead time uncertainty, the safety stock is the best technique in case of the low level of stockout/inventory holding cost, and also in case of a high level of demand variability and low level of lead time variability. On the other hand, safety lead time is the best technique in case of a high level of stockout/inventory holding cost ratio and in case of a high level of demand and lead time variability (Molinder, 1997);

- For the manufacturing/remanufacturing context under stochastic demand, should be adopted different inventory control policies for different PLC phases (introduction, growth, maturity and decline); Moreover, the inventory control policy is not sensitive to the phase length and the demand changing rate (Hsueh, 2011);

- Demand uncertainty is the most considered factor for determining safety stock in different contexts, in contrast, lead-time uncertainty is not sufficiently studied, especially in the MRP environment (M. A. Louly et al., 2008). Table 26 reinforce this statement, showing that demand uncertainty is the most considered uncertainty factor in proposed studies in the literature;

- There are several factors/parameters that should be considered, such as the PLC, demand uncertainty (demand forecast and forecast errors), lead time uncertainty, price fluctuations (e.g., price fluctuation in the market, discount campaigns, promotions), seasonality (sales pattern) and supplier constraints or supply disruptions. Considering only basic parameters (e.g., lead time, actual demand, forecast demand and forecast errors) for calculating safety stock is insufficient. The ERP systems widely used by companies consider values of these parameters from past data to calculate the safety stock using statistical formulas.

In conclusion, the safety stock problem is still an interesting topic for researches and practitioners, since, with the emergence of the Industry 4.0 new challenges have been arisen in all processes of the supply chain. Although the safety stock is used in inventory management to deal with demand and supply uncertainties, it does not solve all problems related to this domain. Other techniques or strategies, such as buffering (reactive) and redesigning (proactive) can also be used for solving or mitigating inventory management problems, such as safety time or capacity buffer. The use of these techniques depends on the specific case study and the context.

4 Supervised learning of supply delay risk

Summary: It is common sense that supply delays have a direct impact on the overall inventory management performance. Typically, delays in orders require to incur in special freights to avoid stockouts and consequently extra costs for the company. Antecipating delays in orders represents enables logistic planners to act proactively, avoiding damages in the production plan. This chapter addresses the prediction of supply delay risk, in which is proposed a machine learning-based approach that aims to help the logistics planners of Bosch AE/P in the decision-making process and consequently improving their efficiency and productivity, as well as avoid extra costs resulting from the occurrence of special freights.

Chapter Table of Contents:

4.1 Advancing Logistics 4.0 with the implementation of a Big Data Warehouse: A Demonstration Case at the Automotive Industry 82

4.1.1 Introduction 82

4.1.2 Related work 84

4.1.3 Propose Architecture for a Logistics 4.0 Big Data Warehouse 86

4.1.3.1 Logical Architecture 86

4.1.3.2 Technological Architecture 88

4.1.4 Demonstration Case 90

4.1.5 Results and Discussion 94

4.1.5.1 Efficacy and Efficiency 94

4.1.5.2 Challenges 96

4.1.5.3 Opportunities 98

4.1.6 Summary and Future Work 100
4.2 A Machine Learning-based framework for predicting supply delay risk using Big Data technology 102
4.2.1 Introduction 103
4.2.2 Aims and contributions 104
4.2.3 Related work 105
4.2.4 Industrial Case Study 107
4.2.5 Material and methods 108
4.2.5.1 Big Data technological architecture 108
4.2.5.2 Machine Learning pipeline 111
4.2.5.3 Data Preparation 111
4.2.5.4 Classification algorithms 116
4.2.5.5 Hyperparameter Tuning 118
4.2.5.6 Evaluation 119
4.2.6 Experiments and Results 123
4.2.6.1 Experimental and modeling setup 123
4.2.6.2 Overall predictive performance 124
4.2.6.3 Misclassification impact on inventory management performance 128
4.2.7 Summary 132

4.1 Advancing Logistics 4.0 with the implementation of a Big Data Warehouse: A Demonstration Case at the Automotive Industry

Nuno Silva[1] · **Júlio Barros**[1] · **Maribel Y. Santos**[2] · **Carlos Costa**[2] · **Paulo Cortez**[2] · **M. Sameiro Carvalho**[3] · **João N.C. Gonçalves**[3]

Electronics **2021**, 10(18), 2221
(https://doi.org/10.3390/electronics10182221)

Abstract The constant advancements in Information Technology have been the main driver of the Big Data concept's success. With it, new concepts like Industry 4.0 and Logistics 4.0 are rising. Due to the increase in data volume, velocity, and variety, organizations are now looking to their data analytics infrastructures and searching for approaches to improve their decision-making capabilities, in order to enhance their results using new approaches such as Big Data and Machine Learning. The implementation of a Big Data Warehouse can be the first step to improve the organizations' data analysis infrastructure and start retrieving value from the usage of Big Data technologies. Moving to Big Data technologies can provide several opportunities for organizations, such as the capability of analysing an enormous quantity of data from different data sources in an efficient way. However, at the same time, different challenges can arise, including data quality, data management, lack of knowledge within the organization, among others. In this work, we propose an approach that can be adopted in the Logistics Department of any organization in order to promote the Logistics 4.0 movement, while highlighting the main challenges and opportunities associated with the development and implementation of a Big Data Warehouse (BDW) in a real demonstration case at a multinational automotive organization.

Keywords Big Data, Data Warehouse, Logistics 4.0, Industry 4.0, Implementation.

4.1.1 Introduction

The explosion of the Information Technologies area has been the driver that launched new concepts such as Big Data and Industry 4.0 into the spotlights. The concept of Industry 4.0 emerged in 2011 from a project created by the German government to promote computerized manufacturing based on new technologies such as AI, Additive Manufacturing (AM), IoT, Big Data (BD), CPS among others (Ghadge et al.,

[1] ALGORITMI Research Centre/LASI, University of Minho, Guimarães 4800–058, Portugal.

[2] ALGORITMI Research Centre/LASI, Department of Information Systems, University of Minho, Guimarães 4800–058, Portugal.

[3] ALGORITMI Research Centre/LASI, Department of Production and Systems, University of Minho, Braga 4710–057, Portugal.

2020; M. Y. Santos, Oliveira e Sá, et al., 2017; C. Tang & Veelenturf, 2019; Winkelhaus & Grosse, 2020). Since the creation of the Industry 4.0 concept that several barriers have hindered its implementation in organizations (even with the evolution of diverse technologies that support it). The financial constraints, the lack of management support, the resistance to change, the lack of infrastructure, and the poor-quality data, among others, are some barriers that need to be faced to implement the concept of Industry 4.0 (Ghadge et al., 2020). This concept relies on the digitization of the production systems to provide the capability of producing customized products within a short time and with costs similar to mass production scenarios (Panetto et al., 2019). This factor has a tremendous impact on the organizations' logistics due to the need of reacting to the sudden changes made by the customers.

The concept of Logistics 4.0 emerged as part of the Industry 4.0 (Kostrzewski et al., 2020), with a few papers being published in recent years (Olesków-Szłapka & Stachowiak, 2019; Strandhagen et al., 2017; Winkelhaus & Grosse, 2020). Logistics 4.0 can be defined as *"... the logistical system that enables the sustainable satisfaction of individualized customer demands without an increase in costs and supports this development in industry and trade using digital technologies"* (Winkelhaus & Grosse, 2020). Such initiative is needed to improve the link between the manufacturers and the customers, in order to avoid failures in the manufacturing system (Winkelhaus & Grosse, 2020).

Throughout history, the evolution suffered by industry also reflected in logistics. In each industrial revolution, a similar evolution occurred in logistics. When the steam power engine was invented and the first industrial revolution appeared, logistics was transformed by using mechanical transport. In the second industrial revolution powered by electricity and mass production, logistics evolved using automatic handling systems. In the third industrial revolution, with the support of information and communications technologies, new logistics management systems were developed (Yavas & Ozkan-Ozen, 2020).

Now, the connection between the concepts of Industry 4.0 and Logistics 4.0 goes deep to the technologies that are used to enforce the Logistics 4.0 main characteristics. Among its characteristics, we can find constant visibility through all supply chain for all stakeholders, verification of the supply chain coherence, and dynamic optimization. These characteristics are enforced by the use of information technologies (Torbacki & Kijewska, 2019).

Big Data technologies, with their capability of analysing massive volumes of diverse data flowing at high velocity, has an important role in the implementation of these new concepts (Industry 4.0 and Logistics 4.0) and in the resolution of their main associated challenges (Strandhagen et al., 2017).

With the implementation of Big Data technologies became possible to perform tasks that involve a massive quantity of data at high speeds such as providing a supply chain control with real-time data, inventory control and management, improving forecasting models, among others (Panetto et al., 2019).

Along with the influence of concepts like Industry 4.0 and Logistics 4.0, the investments in Big Data technologies are being stimulated making them more stable and mature, ready to be implemented inside the organizations and became part of their business.

A vast range of organizations, from diverse types of business, are now trying to evolve their data analyses infrastructures to this new era, advancing their Data Warehouse (DW) based on a more rigid data model to the new concept of BDW with a more dynamic data model (Ngo et al., 2019; Santoso & Yulia, 2017; Sebaa et al., 2018).

This work aims to demonstrate how the implementation of a BDW in a logistics context can drive forward the concept of Logistics 4.0 and improve the organization performance. The contributions of this work are:

- propose a general approach that can be adopted in the Logistics Departments of several organizations;
- propose a logical and technological architecture that supports the BDW and data analysis;
- propose a data model for a logistics BDW, and;
- demonstrate the challenges and opportunities that emerge throughout the development and implementation of a BDW in the Logistics Department.

A demonstration case will be presented, which was developed inside a multinational automotive organization by taking advantage of its existing data platform. The methodology used in this work was the Design Science Research Methodology (DSRM), being this work an outcome of the methodology adoption.

This work is structured as follow: Section 4.1.2 provides the published works related to BDW and their architectures; Section 4.1.3 presents the suggested architecture to solve this problem; Section 4.1.4 describes the organization reality and the tasks performed to accomplish the goal; Section 4.1.5 presents the results accomplished fowled by a discussion where the challenges and opportunities are highlighted; Section 4.1.6 shows the final conclusions and future work.

4.1.2 Related work

With the implementation of concepts like Industry 4.0 and Logistics 4.0, it becomes important to endow the organizations' data analyses infrastructure with the capability of retrieving, transforming and analysing massive amounts of data at high velocity. Before the establishment of the Big Data concept, organizations had their data analyses infrastructure based in DWs where the data model was rigid and structured in order to provide the best performance when data were inquired.

Aftab and Siddiqui (2018) present several differences between a traditional DW and a DW in the era of Big Data. Most of the changes are related to how to deal with data due to their characteristics. Between them, we can highlight few changes such as the change from Extract, Transform and Load (ETL) to Extract, Load and Transform (ELT) that happens to enhance with the processing power of distributed systems, such as Hadoop. The change to real-time and interactive analysis, the change from structured to unstructured data, and the change to analytical interfaces, such as dashboards, based on user requirements.

Nowadays, Big Data technologies, due to their capacity for distributed processing and storage, allow us to have more dynamic data models with less rigid structures, maintaining high performance even with massive volumes of data.

To implement Big Data technologies, we can follow two different approaches: ”the lift and shift”and the ”rip and replace”. ”The lift and shift”strategy means that we replace or extend parts of the existing infrastructure with Big Data technology to improve its capabilities and to solve specific problems. This may result in a use case approach instead of a data-driven approach, which can lead to uncoordinated data silos. The ”rip and replace”approach means that the existing DW is replaced by Big Data technologies (Costa & Santos, 2018).

Independently of these two strategies, there are several architectures and technologies, that can be used to implement a BDW. The use of different types of Not Only SQL (NoSQL) databases, such as document-oriented and column-oriented (Chevalier et al., 2015) or graph models (Gröger et al., 2014) can be used to store the different types of data in the BDW. In the literature, we can find different architectures that can be used in a BDW, such as the Lambda architecture (Kiran et al., 2015) and the NIST Big Data Reference Architecture (NBDRA) (NBD-PWG, 2015). The Lambda architecture has three layers and unifies, in a single software design pattern, the batch and real-time data processing concerns. The three layers presented in the Lambda architecture are batch processing, real-time computing, and a layer to query the data. This division between batch processing and real-time processing allows differentiating data according to their nature and relevance to the business. In this way, it is possible to immediately process the data that is needed in time, while data that is only needed in the long run can be processed later (Kiran et al., 2015).

The NBDRA is presented by its authors as a common reference that can be implemented using any Big Data technology or service provider. It is divided into the following five components: System orchestrator; Data provider; Big Data application provider; Big Data framework provider; and Data consumer. The system orchestrator is the component that establishes the requirements for all the infrastructure, including, among others, architectural design, business requirements, and governance. The data provider is the component that makes data accessible through different interfaces. The Big Data application provider deals with all the necessary tasks to manipulate data through its life cycle. The Big Data framework provider consists of several services or resources that are used by the Big Data application provider. The data consumer is the entity that will take advantage of all the data processing made by the Big Data system (M. Santos & Costa, 2020). Using the NBDRA and the Lambda Architecture as a reference, M. Santos and Costa (2020) created an approach to develop BDWs.

Several examples demonstrate the capacity of Big Data technologies for improving the analytical capabilities of organizations. Chou et al. (2018) propose a system architecture based on Hadoop, Sqoop, Spark, Hive and Impala to analyse data from electrical grids. Sebaa et al. (2018) present an architecture based on the Hadoop ecosystem and a conceptual model to develop a BDW in the Healthcare field. M. Y. Santos, Martinho, and Costa (2017) present a demonstration case where it was applied a Big Data architecture and a set of rules to evolve from a traditional DW to a BDW. Sebaa et al. (2018) developed a BDW based in Hadoop due to its cost-effectiveness, where they present the architecture and the conceptual data model. Ngo et al. (2019), designed and implemented a BDW for agricultural data using Hive, MongoDB and Cassandra. In the same domain, X. Wang et al. (2019) developed and implemented an end to end system for farms management based on Hadoop Distributed File System (HDFS), Spark, Hive and Hbase. Doreswamy et al. (2017) use a hybrid DW model with OLTP a system and Hadoop to develop a meteorological DW using a star schema. Costa and Santos (2017b) developed a BDW for smart cities using technologies such as Hive, Cassandra, HDFS, Presto, among others. Vieira et al. (2020) developed a tool using Big Data technologies and a simulation model to assess the impact of disruptions in the performance of the supply chain.

These examples demonstrate how Big Data technologies can be used in collaboration with traditional DW or even replacing them, both aiming to improve the analytical capabilities of the organizations.

Although several domains are addressed in the literature, the lack of work in the logistics area is notorious. Moreover, few approach the problems faced when the implementation occurs in the real world.

4.1.3 Propose Architecture for a Logistics 4.0 Big Data Warehouse

In this section, it is presented the logical (4.1.3.1) and technological (4.1.3.2) architectures that can be used to implement a BDW for the Logistics 4.0 movement.

4.1.3.1 Logical Architecture

The main goal of this BDW is to be an analytical repository containing a substantial amount of data, in order to support the daily activities of the logistics decision-makers in the Logistics 4.0 era.

Two of the key factors in Logistics 4.0 are the real-time exchange of information between all the actors in the supply chain and the real-time Big Data analytics of vehicles, products and facilities location (Strandhagen et al., 2017).

The exchange of information between all actors in the supply chain can originate diverse data sources with different types of data that need to be stored and analysed in one central repository in order to be easily accessible by practitioners. The same happens with the real-time BDA of the diverse supply chain components (vehicles, products, and facilities location). Considering this, the real-time characteristics can be important, nevertheless, it is necessary to adapt to the organizational requirements. Real-time analytics can be a different concept from one organization to other. For example, for one organization, the requirements of real-time can be to have access to data in less than ten seconds, but for other organizations it can be to access the data in less than two minutes. Moreover, some organizations do not need to create an architecture that takes into consideration the real-time requirements.

In our demonstration case, the organization does not have the requirement of real-time analysis, so the architecture presented in Figure 35 does not incorporate that component. Nevertheless, due to the relevance of real-time in Logistics 4.0, it may be relevant to implement and validate that component in future work.

As can be seen in Figure 35, the logical architecture has the following components:

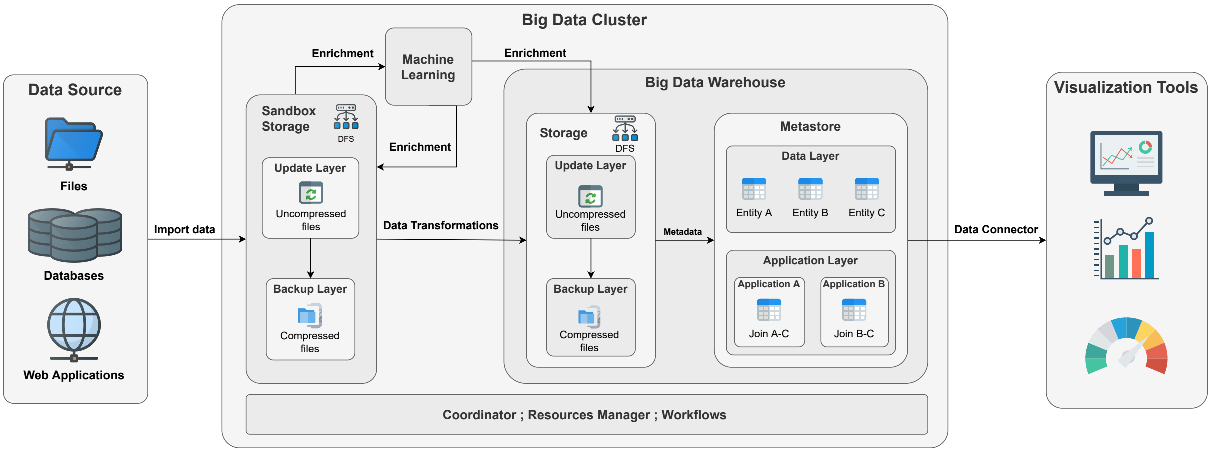


Figure 35: Logical architecture.

- **Sandbox Storage:** where the raw data is stored in a distributed file system before any transformation. This component is divided into two layers: Update Layer and Backup Layer. The Update Layer contains the up-to-date data retrieved from the sources, while the Backup Layer contains compressed outdated data to be used in case of necessity.

- **BDW Storage:** where data is stored in the distributed file system and accessible using the metastore after being transformed. This component has two layers with the same functionality as the Sandbox Storage layers: i) a layer that provides updated data, ii) and another layer to provide a backup in case of problems with the new data.

- **Machine Learning component:** uses raw data from the Sandbox storage or clean data from the BDW to create predictions, in order to enrich the data and store it in the Sandbox Storage or in the BDW to provide predictive capabilities for the organization. This component can increase the organization's capabilities to understand and predict changes in their supply chain and be capable to adapt quickly.

- **Metastore:** provides an interface to access the stored data. This component is divided into two layers: i) the data layer where the data is modelled using a data-driven approach, and; ii) the application layer where we have the necessary materialized objects or views to answer the needs of specific applications. The existence of these two layers provides some advantages. One of these advantages is the capability of creating several abstractions on top of the data layer, providing a simple and fast way to access the data. In this application layer, each application can have its views or tables (materialized objects), increasing the performance when accessing the data. Moreover, if the organization has different teams working in different applications, if necessary, each team can create the necessary tables or views for their application, providing higher business agility.

- **The Coordinator, Resources Management and Workflows:** provide functionalities to manage the BD Cluster and the data life cycle. The Coordinator and Workflow allow the creation of diverse jobs or tasks that can be submitted in the desired order. The Resource Manager distributes the clusters resources to process the jobs.

Outside the BD Cluster, we can find the data sources that provide the raw data to be used in the BDW and the Visualizations Tools where dashboards are developed to present the results to the users.

4.1.3.2 Technological Architecture

Due to the need of analysing big quantities of data in the most efficient way, new technologies that use the power of distributed processing and storage have gained significant attention. Probably the most well-known technology in this context, which can arguably be seen as the originating driver of the Big Data movement, is Apache Hadoop, where data can be stored in the HDFS (Shvachko et al., 2010) and then processed using the Map and Reduce (Dean & Ghemawat, 2008) programming model. Several other technologies such as Sqoop, Hive (Thusoo et al., 2009), Spark (Spark, 2018), and Impala (Bittorf et al., 2015), among others, are being constantly developed to tackle specific problems in the Big Data ecosystem. These technologies allow the practitioners to retrieve data from the data sources, store it with appropriate metadata and then processing it, in order to provide useful knowledge to the end-users.

Currently, in the BD world, the amount of BD technologies is overwhelming and sometimes can be difficult to understand and choose the right technology for the right job. For example, for data collection, technologies such as Flume, Kafka, or Talend can be used. For data preparation and enrichment, we can use Spark or Storm. For data storage, Hive with HDFS, NoSQL databases, or Kudu can be used. For machine learning tasks, we can use Spark, H2O, and TensorFlow L'Heureux et al. (2017). For query engines, Impala, Presto, or Drill can be used. For data visualization, tools like Tableau, Power BI, JavaScript can be used (Costa et al., 2018).

Due to the organizational requirements and due to the technologies available in the organization depicted in this demonstration case, the technological architecture presented in Figure 36 was used to support this demonstration case. Nerveless this technological architecture can be used inside others organization's Logistics Departments, assuming the goals and requirements are similar to the ones depicted in this work. In case of distinct requirements, some technologies could be adjusted. Regarding data ingestion from the sources, this work uses Sqoop. Even though Sqoop can only connect to structured databases (Aravinth et al., 2015), since for this demonstration case the organization's data sources were only Structured Query Language (SQL) databases, there was no need to use another technology to ingest the data. After the data is retrieved from the sources, the same is stored in HDFS, using the Parquet format, which is one of the several formats that can be used to store data in HDFS. Other formats that can be used are, for example, ORC or AVRO (Ivanov & Pergolesi, 2020). Parquet was chosen not only due to its adequate compatibility with Spark and Impala technology but also due to its read-oriented format and with adequate compression, which will bring advantages when we need to query the data (Baranowski et al., 2015). Moreover, it was necessary to develop a Bash script in order to provide a mechanism to create data backups in the Sandbox Storage and in the BDW.

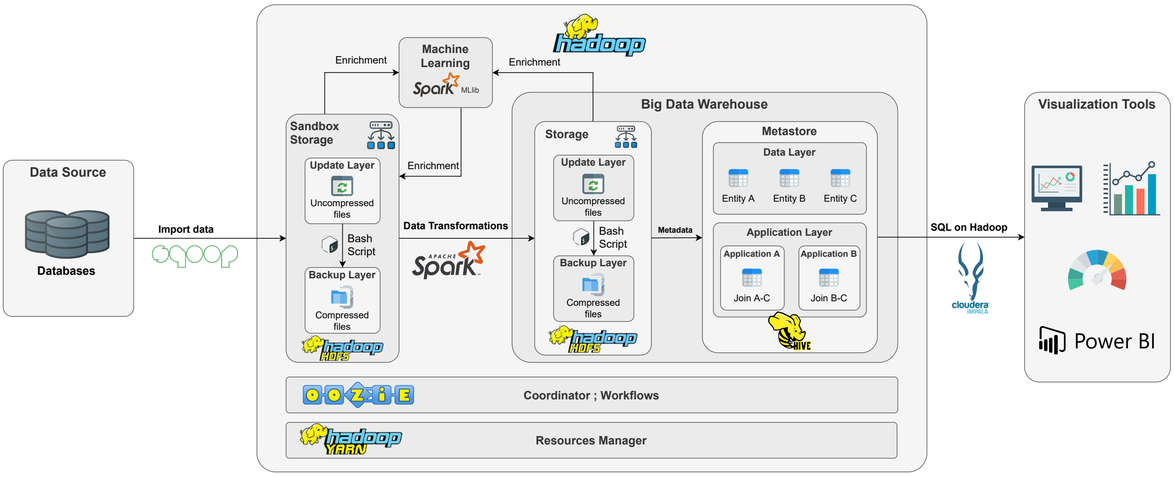


Figure 36: Technological architecture.

Spark was the chosen framework due to its data cleansing and transformation capabilities and due to the capability to develop several machine learning models. Spark has the SparkSQL (Armbrust et al., 2015) library that allows the use of SQL functions in conjunction with the Spark programming Application Programming Interface (API) and complex libraries such as Spark MlLib (Meng et al., 2016). Being able to perform all these tasks in one unique framework is a significant advantage, since, in this way, it is not necessary to spend more time using and learning different technologies. Moreover, Spark is compatible with Parquet files and Hive, which will be used to provide the data and metadata to the end-users.

Hive includes the Hive Metastore (the system catalog) where the metadata (schema and other statistics) are stored, allowing proper data exploration and query optimizations (Thusoo et al., 2009). Hive allows the creation of external tables where data is stored in HDFS directories and its life cycle is not managed by Hive (Thusoo et al., 2009). Within Hive, we create two levels of interaction with the data. In the first level, the data is modelled using a data-driven approach where the core entities (such as Needs, Stocks, Products, among others) and other entities like Date and Time are stored. This layer allows ad hoc access to the data from these entities to be used by any team or project. In the second layer, the application layer, a new set of objects (materialized tables or views), oriented to the applications' needs, are created to provide access to the specific data that each application or project needs. This will provide more personalized access to the data that will increase the application performance and higher business agility, thus each team can create their tables or views as they need.

Impala provides a Massively Parallel Processing (MPP) SQL engine that combines the flexibility and scalability of Hadoop with the familiarity of SQL and has proven to be generally faster than Spark or Hive according to Qin et al. (2017) and to Bittorf et al. (2015). Impala can too be used to query data from HBase and provide a connection to visualization applications, such as Tableau or Power BI, where dashboards can be developed to present to the end-user the knowledge retrieved from the data (Bittorf et al., 2015).

This technological architecture supports all the requirements of this project, granting that we can allow the data analysis team to provide knowledge to be used by the end-users, in order to support their decisions and therefore improving the organization's results. Moreover, it can be used in other Logistics 4.0 projects to create a new centralized repository that aggregates different data sources and requires predictive capabilities.

4.1.4 Demonstration Case

The application domain addressed in this paper is the Logistics Innovation Department of an automotive factory. In this context, the Logistics Department handles large volumes of data related to nearly 7000 raw materials from a set of about 400 suppliers spread all over the globe, which impact the production of about 1100 finished products. Concerning to internal logistics management, the department is responsible for monitoring and analysing data and material movements referring to approximately 85 daily scheduled deliveries, in order to ensure the supply of material necessary for the proper functioning of about 100 production lines associated with various high-service level customers. In light of the complexity of the organization's supply chain topology, the organization intends to foster the proposal, development and evaluation of BDA tools capable of integrating and automating a large part of the logistics processes that, until now, are managed by conventional spreadsheets extracted from classic and parameterizable MRP methodologies existing in a given ERP system.

It is an essential department inside of a production facility and deals on a daily basis with orders, deliveries, delays, production plans, inventory, among other processes. These business processes are crucial to maintain the production lines working and to deliver in time the finished goods to the clients. It is a complex and enormous department with countless business processes.

Due to this complexity, the implementation of a BDW needs to be addressed in an interactive way, choosing one process at a time, looking at the data sources, selecting the appropriated attributes and modeling the data in a data-driven approach that has as a final goal an integrated BDW supporting Logistics 4.0.

Therefore, in this specific case, to start the BDW proposal we analysed the processes that should be considered the core component of this BDW. With the collaboration of key experts in the Logistics Department, the following processes were selected: Product Inventory, Delivery, Purchase Order, and Needs. This is the first task in the development process presented in Figure 37.

These processes will be the main drivers of the analytical objects in the BDW. Besides these objects, other objects will be created, such as a spatial object with information related to countries, Date and Time objects, and complementary analytical objects such as Product, Plant and Vendor. Each one of these processes is supported by one or more tables in the ERP used by the organization. These different types of objects are explained later in this section.

The understanding and selection of the business processes, together with the understanding and selection of the data sources, compose the first activity of the development process (Figure 37) called Data Understanding. In this activity, it is necessary to understand the data from the data sources, namely the tables associated with each business process, how they are related, their private and foreign keys, the meaning and possible values of each attribute, among other steps. The second task is to select what tables will be used to develop the BDW.

The next activity is related to the "Data Quality"activity. Data quality is one of the most important tasks in data-related projects. In this case, this activity has significant importance due to the complexity of the data sources and their high number of attributes. For example, some transactional tables have more than 200 attributes, although many of them are not used. In our demonstration case, data quality criteria were defined to verify if an attribute will be used in the BDW. In this specific case, we established that any attribute with more than 90% of empty or nulls values will not be used. This rule was essential to limit the number of used attributes, excluding the ones that have low analytical value. Another rule that was used was to manually verify if the attributes with only one or two distinct values were worth using. All these rules were defined considering the organizational and decision-making context. The next step was to produce the data quality reports through the execution of several spark jobs that analysed the data extracted from HDFS. The attributes that will be part of the BDW are selected applying the previously defined data quality criteria.

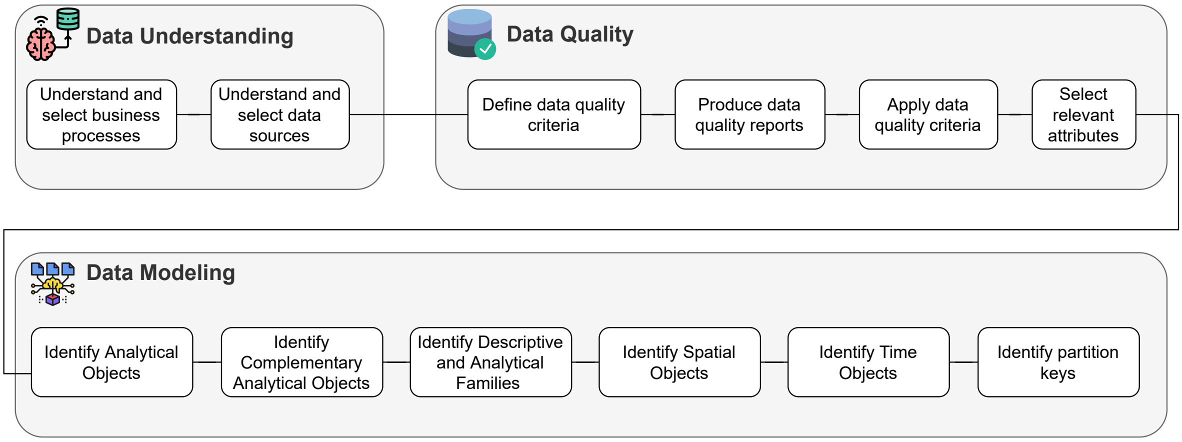

Data Understanding
Understand and select business processes
Understand and select data sources
Data Quality
Define data quality criteria
Produce data quality reports
Apply data quality criteria
Select relevant attributes
Data Modeling
Identify Analytical Objects
Identify Complementary Analytical Objects
Identify Descriptive and Analytical Families
Identify Spatial Objects
Identify Time Objects
Identify partition keys

Figure 37: Development process.

After the Data Understanding and the Data Quality, it was possible to model the BDW. To do that, the modeling methodology presented by Santos and Costa M. Santos and Costa (2020) was applied in order to propose a data model capable of integrating a significant amount of data. The methodology is based on the creation of the following objects: Analytical Objects, Complementary Analytical Objects, Spatial Objects, Time Object, and Date Object.

An Analytical Object is a subject of interest, highly denormalized and that can answer queries by itself avoiding joins with other objects. These objects are directly related to the business processes such as sales or deliveries and should be the firsts to be analysed and identified in order to verify if it is necessary, or not, to create Complementary Analytical Objects. A Complementary Analytical Object is an object that includes attributes usually used or shared by different Analytical Objects and that can be used to complement the analysis of other objects, such as the Analytical Objects. Each object can be divided into two distinct parts, the descriptive and analytical families. These families provide a logical group for the object attributes depending on their type and purpose. The descriptive family group all the attributes that can provide different perspectives of analysis of the business indicators, while the analytical family group the attributes with those business indicators to analyse the business process or part of it. These objects can be integrated with the use of join operations (M. Santos & Costa, 2020). Figure 38 presents the data model identified with the application of this methodology. Due to privacy concerns, it is only possible to disclose some of the attributes present in the several objects This data model was developed in the logistics context of this specific factory but can be used as starting point for any Logistics Department of any organization.

The Analytical Objects used in this work are: Product Inventory that has all information about the stocks of each product; Deliveries that has information about when each order is delivered; Purchase Order that has information about how many products are ordered; and Needs that has information about production lines needs.

The Time and Date objects were created from scratch and populated with information related to each one. For example, in the Date object, we created boolean attributes such as week_day, weekend, summer, winter, monday, tuesday, and others. In the Time object, attributes such as lunch-time, in-office, out-office, rush hour, were created. This allowed us to analyse the relevant information and contextualize it in time and date.

The Complementary Analytical Objects had emerged in the data modeling process due to the need of analysing different Analytical Objects using data from the Complementary Analytical Objects. In these objects was stored relevant and specific data that can provide useful information when used together with data from several Analytical Objects. From these objects, we can highlight the following: Plant, Product, and Vendor.

The object Country is a Spatial Object due to the geographical domain that includes information from the transactional database and from a JavaScript Object Notation (JSON) file (already stored in HDFS) with more information, such as the continent name.

The implementation process presented in Figure 39 starts with the data extraction performed using Sqoop and Oozie Workflows and all data was stored in a HDFS directory called Sandbox. This Sandbox directory allows the storage of all raw data and it is divided into sub-directories where each data source has its own directory and is divided into tables or entities. In this demonstration case, two data sources were used, the transactional database and a JSON file.

With all the necessary data stored in HDFS, we can use Spark to perform the data transformation phase, where transformations and partitions keys are identified. Moreover, it is in this phase that the data enrichment can be performed with predictions from the machine learning models.

After the data transformation, the data is stored in the BDW where each table represents one of the objects included in the data model. Moreover, when the size of the object is too large to be used as one unique file, the object is partitioned according to its partition keys in order to improve the performance when querying the data. Furthermore, external Hive tables were created to provide Impala access to data. Impala will be the SQL query engine that allows the connection between Power BI and the data stored in

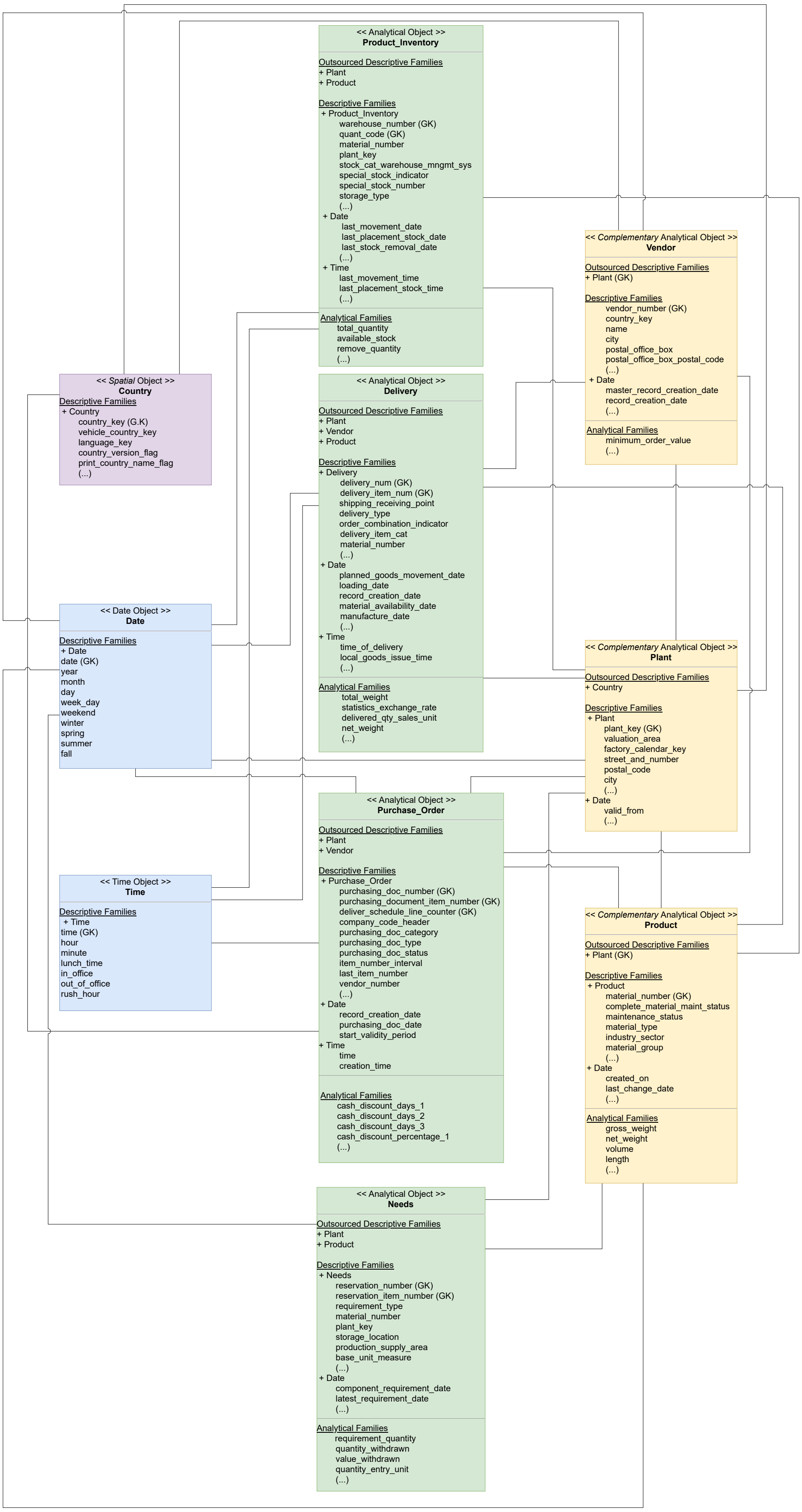


Figure 38: BDW Data Model.

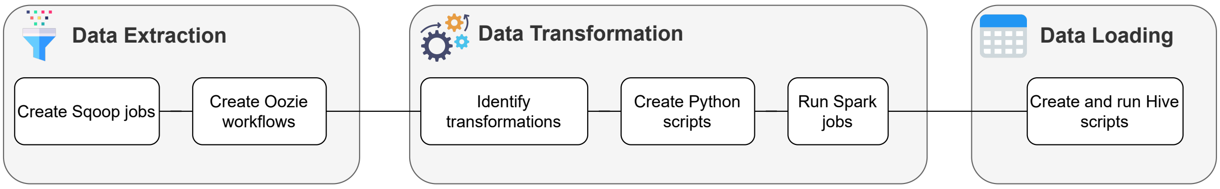


Figure 39: Implementation Process.

HDFS.

4.1.5 Results and Discussion

In this section, we discuss the efficacy and efficiency (4.1.5.1) of the BDW implementation. In subsections 4.1.5.2 and 4.1.5.3, the challenges and opportunities faced in the development of this work are present.

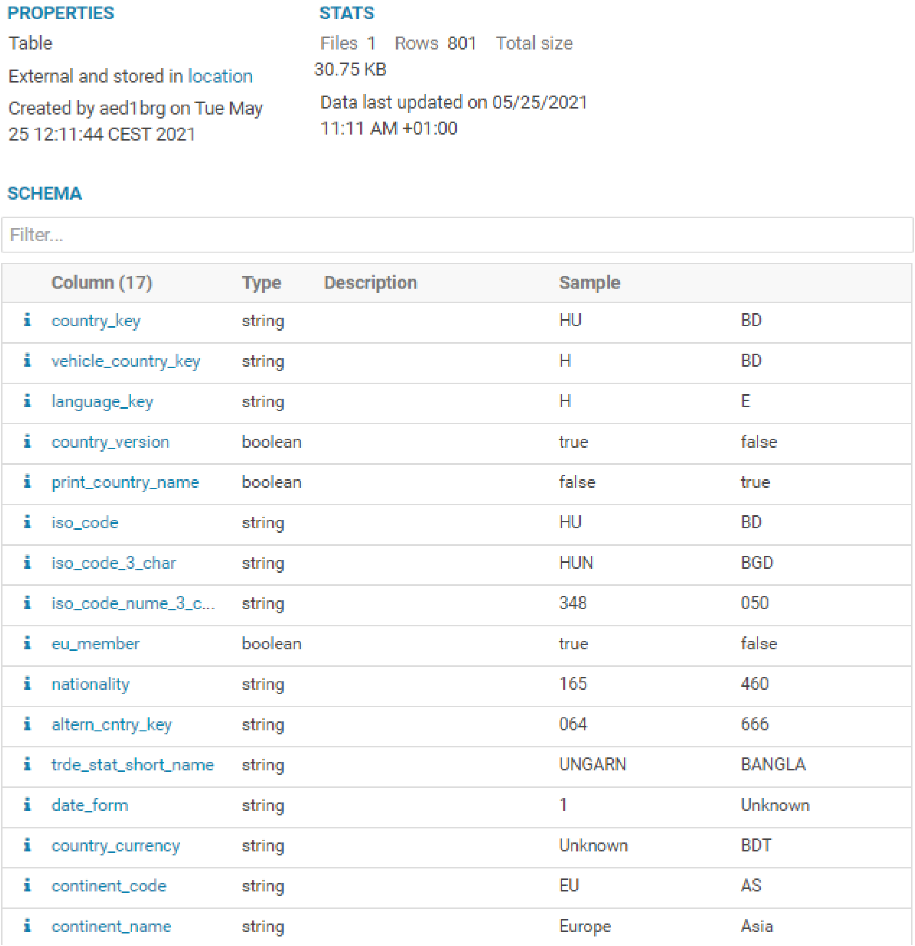
4.1.5.1 Efficacy and Efficiency

With the BDW implementation, it was possible to create a data repository that includes several businesses processes of the Logistics Department. Each process contains data from one or more tables from the transactional database used by the organization.

The data model is dynamic and able to change quickly, in order to include more tables, with more information related to any object that already exists in the BDW or to create new ones. The Time and Date objects can be used with other objects to understand the organization temporal dynamics, such as understand if there are any specific moments in the year where more delays are verified, or even when the suppliers are usually late with the deliveries. Similar reasoning can be used with the objects Plant and Inventory to analyse which plant has more inventory in its storage facilities.

With this work, it is now possible for the practitioners to use raw data extracted from the data sources (using the Sandbox layer) or use data already cleaned and transformed using the BDW layer. This can be achieved using the BDW Hive tables (as an example, Figure 40 shows the Country table view using the HUE interface) or the parquet files stored in the HDFS. They can also create specific materialized objects in the Application Layer in order to decrease the time needed to query the data. This reduces or even avoid the initial development time needed to understand, extract, store, and transform data.

The Machine Learning component can also use data from the different architecture components to provide useful predictions. For example, the available data can be used to predict if some scheduled delivery will be late or not. With this information, the logistics planners can take several actions to reduce the impact of this situation. This can be achieved using data from the Sandbox or from the BDW. Machine Learning models can be created with this data using the Spark ML framework. Both the model and the predictions are stored in the HDFS being available for later use and for possible updates in the future. Furthermore, this data is now accessible to the organization through Impala connector and can be used to provide different insights about the organization status, or even in projects that use ML to predict or



PROPERTIES

Table

External and stored in location

Created by aed1brg on Tue May 25 12:11:44 CEST 2021

STATS

Files 1 Rows 801 Total size 30.75 KB

Data last updated on 05/25/2021 11:11 AM +01:00

SCHEMA

Filter...

	Column (17)	Type	Description	Sample	
ℹ	country_key	string		HU	BD
ℹ	vehicle_country_key	string		H	BD
ℹ	language_key	string		H	E
ℹ	country_version	boolean		true	false
ℹ	print_country_name	boolean		false	true
ℹ	iso_code	string		HU	BD
ℹ	iso_code_3_char	string		HUN	BGD
ℹ	iso_code_nume_3_c...	string		348	050
ℹ	eu_member	boolean		true	false
ℹ	nationality	string		165	460
ℹ	altern_cntry_key	string		064	666
ℹ	trde_stat_short_name	string		UNGARN	BANGLA
ℹ	date_form	string		1	Unknown
ℹ	country_currency	string		Unknown	BDT
ℹ	continent_code	string		EU	AS
ℹ	continent_name	string		Europe	Asia

Figure 40: Country table in Hive.

classify data to help in the decision making. This means that the time and the necessary knowledge to develop useful dashboards for management is smaller. In Figure 41, a dashboard that analyses historical and predicted data is present, showing information about deliveries. It is an overview where the historical and predicted delayed or at time deliveries are analysed in several dimensions.

The top right component of the dashboard shows the number of products that belongs to each category (A, B or C). This product classification demonstrates how important is each product for the organization. Products classified with “A” mean that these are expensive products for the organization and normally with more lead time, for example, electronic screens. The “B” category is for less expensive products, and the “C” category is for cheap products such as bolts. The impact on delays for products classified with “A” is superior to the products classified with “B” and “C”. The graph shows that there is a bigger number of deliveries of “C” classification products demonstrating that this type of product has more frequent deliveries. So, if for some reason there is a shortage in stock of this product type, the organization will be able to solve that problem rapidly.

The two graphs in the lower-left corner of the dashboard compare the on-time deliveries and the delayed deliveries analysed by the season year. Each one compares the historical data and the predictions made by the machine learning algorithm. The left one shows that the predictions followed the trend of the historical date. The right one shows that an increase in delays in Autumn is predicted. With this information, the organization can prepare mitigation actions to decrease the impact of the delays.

The middle graphs compare the delayed deliveries and on-time deliveries by transportation mode. For example, we can see that the predictions (centre lower graph) show a general increase in the percentage of on-time deliveries.

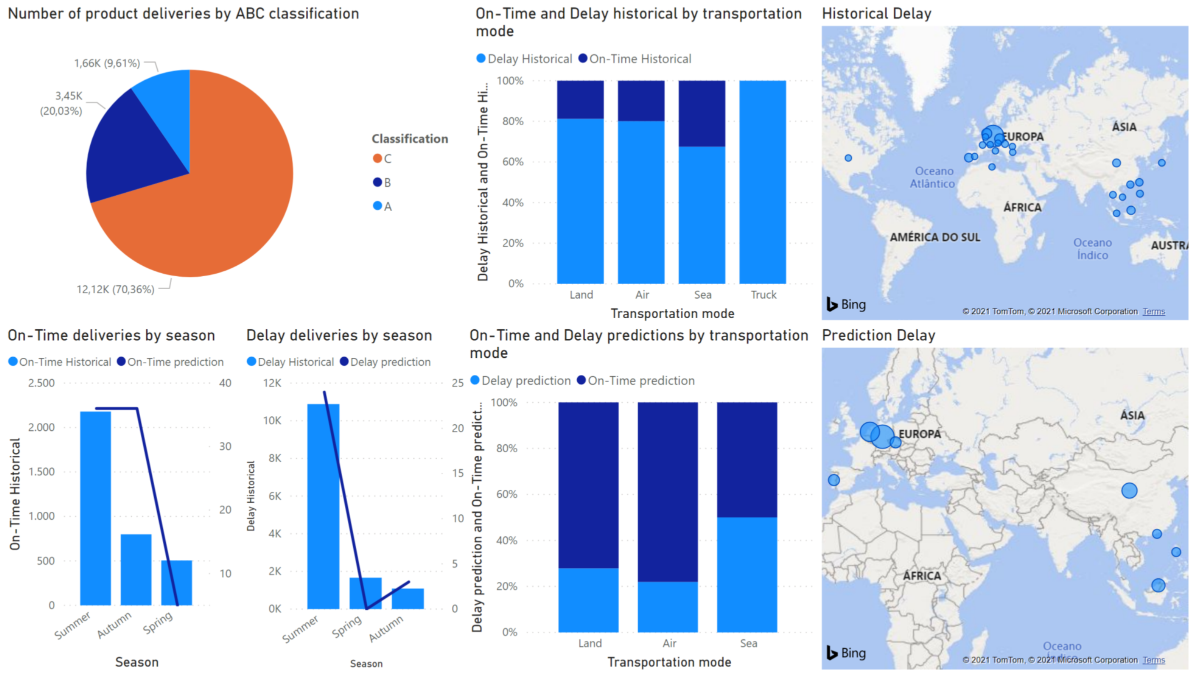

Number of product deliveries by ABC classification
1,66K (9,61%)
3,45K (20,03%)
12,12K (70,36%)
Classification
● C
● B
● A
On-Time and Delay historical by transportation mode
● Delay Historical ● On-Time Historical
Delay Historical and On-Time Hi...
100%
80%
60%
40%
20%
0%
Land
Air
Sea
Truck
Transportation mode
Historical Delay
EUROPA
ÁSIA
Oceano Atlântico
ÁFRICA
AMÉRICA DO SUL
Oceano Índico
AUSTRA
Bing
© 2021 TomTom, © 2021 Microsoft Corporation Terms
On-Time deliveries by season
● On-Time Historical ● On-Time prediction
On-Time Historical
2.500
2.000
1.500
1.000
500
0
40
30
20
10
Summer
Autumn
Spring
Season
Delay deliveries by season
● Delay Historical ● Delay prediction
Delay Historical
12K
10K
8K
6K
4K
2K
0K
25
20
15
10
5
0
Summer
Spring
Autumn
Season
On-Time and Delay predictions by transportation mode
● Delay prediction ● On-Time prediction
Delay prediction and On-Time predict...
100%
80%
60%
40%
20%
0%
Land
Air
Sea
Transportation mode
Prediction Delay
EUROPA
ÁSIA
ÁFRICA
Oceano
Bing
© 2021 TomTom, © 2021 Microsoft Corporation Terms

Figure 41: Dashboard with historical and predicted data related with deliveries.

The right side graphs compare the historical data with delays and the predictions. Bigger circles point that are more deliveries from those countries will arrive with delays. We can see more delays from products shipped by European countries. The same is predicted by the machine learning algorithm.

These results are based on a portion of the historical data provided by the organization. In future work, the accuracy of the predictions will be verified to see if they conform with the organization's needs. More data will be also used to improve the model quality.

4.1.5.2 Challenges

The implementation of new technology inside the organization's Logistics Department can be difficult and rises diverse types of challenges. These challenges can be related to the technology itself, with the lack of knowledge to develop the project, with the organizational culture, with the time and the cost to develop the project, among others. When that technology will use or rely on the provided transactional data to be successful, several new types of challenges related to data emerge.

Moreover, if the organization has a large dimension, can be extremely difficult to get the necessary knowledge to understand the different business processes inside the Logistics Department and the data generated by them. For example, if we are inside of a multinational organization, with diverse divisions, spread by multiple countries, with a complex transactional database, the data understanding will be one of the most challenging steps in the project.

The following list provides the identification and brief characterization of the most relevant challenges that were faced through the development of this work.

A. **Data and technological challenges**

- **Data Understanding**
  Understanding the data that is stored in the transactional database is usually a challenge, even worse when the organization is a multinational with a considerable dimension. Transactional databases are complex systems, with misleading tables and attributes names. The existing documentation about the data source is usually sparse, not given enough insights about the data. Several logistics concepts need to be known, such as safety stock, safety time, delivery time, procurement, among others, in order to better understand the data and their relationships.
- **Poor or missing raw data**
  When an organization starts a project that will use the raw data generated by the daily business, it is necessary to identify if the necessary data is being generated and stored in the transactional system and its overall quality. Sometimes the project goals can not be achieved due to the lack of data or data with quality. In complex ERP systems is possible to verify that many attributes are not used by the organization. For example in logistics, knowing where an order is in transit to its destination can be very useful to predict if it will be on time, or not, and to make decisions about how to avoid stops in the production line.
- **Different values in different data sources for the same attribute**
  Due to the large and complex transactional system, is fairly common to find the same attribute in different tables, related to the same entity, but with different values. Understand why this happens and understand the type of situations that motivate this type of behaviour can be difficult.
- **Technological infrastructure**
  The adequate technological infrastructure is essential to stable a project development. In an organization, the technological infrastructures can be based on outdated technology or the technological infrastructure can change during the project lifetime. This will lead to a project adaptation to the existing technologies or their evolution as the infrastructure change.

B. **Organizational challenges**

- **Access to data and to a technological infrastructure**
  One of the first tasks in projects of this nature is to get access to data and to the infrastructure

that will be used to process and store it. This is a task that needs to be done at the beginning of the project and where the organizations' policies can interfere in a negative way. This can not be an obstacle or take a long time to overcome.

- **Understand the business processes**
  Commonly, large organizations have many and complex business processes, with diverse rules, exceptions and paths, which can be difficult to understand. Moreover, the documentation about the business processes can be insufficient, creating another obstacle in this type of project. In the logistics area, where daily interactions with the suppliers and their systems exist, where processes are complex in order to achieve better results in the production line, and where concepts such as just in time production are being implemented, the documentations has a relevant impact when new projects start to be developed.

C. **Project team challenges**

- **Lack of knowledge in the used technologies**
  As Big Data is a recent concept, there is a lack of human resources with experience in the technologies used to support this concept. Building a team without any experience in Big Data can lead to several problems in the project. Moreover, when adding specific requirements of a complex area like logistics, more difficult is to get multidisciplinary teams with knowledge in both areas.
- **Lack of sufficient human resources**
  To develop such a complex project, the project team needs an adequate number of human resources. The lack of sufficient human resources can cause delays in project development. Teams with a high number of elements can be prejudicial to the project too, but very small teams lead to a lack of different backgrounds and points of view that can hinder the project.

The challenges enumerated in this section are some of the biggest challenges that a team can encounter while develop and implement a BDW inside of an organization with a considerable size. The challenges can cause delays in the project milestones and they should be taken into account when the project is planned. Most of them can be mitigated with simple actions such as grant early access to all necessary resources and develop the necessary documentation in all projects.

4.1.5.3 Opportunities

When an organization go through a technological change such as the creation of a BDW, some opportunities emerge. Indeed, we can say that each challenge can be transformed into one opportunity. Therefore, we will take the challenges provided in section 4.1.5.2 and transform them into opportunities.

A. **Data and technological opportunities**

- **Improve documentation**
  Very often, documentation is treated as the less important part of the project. The time and effort put in the documentation development are lower than required, leading to poor documentation. With the development of a new project, the poor documentation of the previous one becomes evident. The effort that needs to be done to understand the previous project can be reused to improve the documentation and, therefore, decrease the time and effort needed for the next ones.
- **Improve data quality**
  Data quality is essential to the development of these data-based projects. As we need to perform data quality tasks, this can be used to detect and report data problems that can be fixed in the near future. This can be useful not only for this project but even for past and future projects.
- **Technological infrastructure**
  A new project that requires new technology can be an excellent driver to improve the technological infrastructure existent in the organization. These changes can include, for example, updating the existent technologies or the implementation of new ones.

B. **Organizational opportunities**

- **Improve internal processes**
  With the implementation of new technology, some internal processes will be analysed and can be improved. Moreover, processes can use the newly available technology to improve their performance.
- **Improve business processes documentation**
  Many analytical teams do not know the business processes and they need to found the right person to ask. Often, if they ask the same question to different persons, they will get different answers. Properly document the business processes can be a key way to improve the business understanding not only inside the analytical teams but for the organization in general.

C. **Project team opportunities**

- **Creation of a team specialized in Big Data technologies**
  Research projects can have a tremendous impact on organizations, not only by the obtained results but also by the improved capabilities of human resources. In this specific case, the creation of one team specialized in Big Data technologies can boost more projects, more efficiently, and with more efficacy.

- **Improve workers knowledge in logistics processes**
  Human resources with more business knowledge can bring their knowledge to other projects and have a positive impact on them. This can be verified not only in new ones but also in the maintenance and improvement of other ongoing projects.
- **Improve workers knowledge about data sources**
  Data analytics projects always depend on the data source. Knowledge about them is essential for a good start and a proper development of the project. It is crucial to have in the project team, at least, one specialized resource in the data sources, helping the development team to understand the data.

Besides the enumerated opportunities, other opportunities can arise with the creation and implementation of a BDW in a Logistics Department. For example, new projects can be initiated and use the BDW as their data source, providing integrated and consolidated data for their timely development. Other departments can use data in the BDW to improve their predictions and their decision making needs.

4.1.6 Summary and Future Work

This paper presented the proposal and implementation of a BDW into a Logistics Department of an automotive factory. The implementation of the BDW is the starting point to push the concept of Logistics 4.0 in this facility, improving the analytical capabilities and supporting the decision-making process in the Logistics Department. Moreover, we highlight several challenges and opportunities that normally are not considered in other works.

Through this work, we presented the logical and technological architecture that support the implementation of the BDW that includes several logistics processes. Moreover, we presented the proposed BDW data model. The BDW data model is a key element to get insights into the current state of the organization and to support the logistics planners' decisions efficiently. The logical and technological architecture, as well as the data model, can be used as starting a point to develop and implement a BDW in similar Logistics Departments.

As we advance, we faced several challenges and opportunities in the BDW development and implementation. One of the most difficult challenges was to understand the several logistics processes and how the data of these processes is stored in the transactional system. Finding the right data to support the proposed system was a difficult and time-consuming task. Nevertheless, the most important thing is to be aware of the challenges and implement mitigation plans in order to solve them, or at least decrease their impact on the project final results. Other challenges that can be faced in this area are related to the technologies and the available infrastructure used by the organization. Sometimes the technological infrastructure is changing during the project what can lead to several project changes. Moreover, the available infrastructure can include outdated technologies or be short in resources when used by several teams at the same time.

In the opportunities field, several points that can be addressed to improve the organization, the Logistics Department, and the next projects. But these opportunities need to be addressed in new projects with a well-defined goal and scope, due to the new challenges that these projects will rise. Organizations need to promote a culture of continuous improvement to face these opportunities.

As future work, the BDW implementation can be improved by automatizing the data extraction, transforming, and enrichment pipelines to increase the performance and decrease the human intervention. Moreover, the data model can be extended by adding new objects (complementary or analytical) in order to enlarge their scope or improving the existent ones by adding new data to the already existing objects. Furthermore, more machine learning models can be created and integrated into the existing BDW to enrich the data and provide predictions to help the logistics planners. Also, the implementation of a real-time layer should be taken into consideration.

4.2 A Machine Learning-based framework for predicting supply delay risk using Big Data technology

Júlio Barros[1] · **Nuno Silva**[1] · **João N.C. Gonçalves**[2] · **Paulo Cortez**[3] · **M. Sameiro Carvalho**[2] · **Maribel Y. Santos**[3] · **Carlos Costa**[3]

Submitted to a Journal

Abstract

In supply chain (SC) management, Big Data Analytics (BDA) has been explored by researchers and practitioners as a way to manage uncertainty factors and try to mitigate their effects on SC performance. Focusing on the upstream SC, it is well-known that supply delays impact strongly on inventory and demand management. However, while optimization-based techniques have been the most popular operations research approach to tackle SC risk management related problems, only recently have we witnessed the application of Machine Learning (ML) techniques in the upstream SC side, naturally subject to uncertainty signals emerging from downstream SC players. In this paper, we propose a novel ML-based framework for predicting the risk of supply delay using a scalable technological BDA architecture. Our main contribution relates to the introduction of a framework that combines ML and BDA to predict the risk of supply delays while evaluating the financial impact of model misclassifications. The proposed ML approach includes a new set of data features whose dynamics may affect positively or negatively the supplier delivery performance. Moreover, to the best of our knowledge, this work is the first in the SC management literature that combines BDA and ML in the context of supply risk identification. In sharp contrast with common practice, where theoretical frameworks are typically adopted, our approach takes advantage of real-world empirical data to evaluate each classifier, under a rolling window validation scheme, in terms of both predictive power and the potential impact of misclassification on the overall inventory management performance. When comparing the performance of six ML classifiers, our results favor the Random Forest (Random Forest (RF)) algorithm, which provides an excellent predictive discrimination power (90%) and the lowest misclassification costs (average of 1.3k monetary units). .

Keywords On-time delivery, Supplier delay, Supply chain risks, Machine Learning, Big Data Analytics.

[1] ALGORITMI Research Centre/LASI, University of Minho, Guimarães 4800–058, Portugal.

[2] ALGORITMI Research Centre/LASI, Department of Production and Systems, University of Minho, Braga 4710–057, Portugal.

[3] ALGORITMI Research Centre/LASI, Department of Information Systems, University of Minho, Guimarães 4800–058, Portugal.

4.2.1 Introduction

Industry 4.0 represents the adoption of Information Technology (IT) into the industry to achieve a higher level of operational efficiency, productivity, and automation (Drath & Horch, 2014; Manavalan & Jayakrishna, 2019; H. Xu et al., 2018; L. Yang, 2017). Key concepts such as IoT and CPS emerge with Industry 4.0, transforming traditional factories into smart factories. These factories of the future or smart factories produce a massive amount of industrial data that can be converted in valuable insights for the company through data analysis and integration (Božič & Dimovski, 2019; Costa & Santos, 2017a; Govindan et al., 2018; Trkman et al., 2010; Waller & Fawcett, 2013). In this context, BDA can be used to improve the decision-making process, create value and competitive advantages to companies (Božič & Dimovski, 2019; Roßmann et al., 2018; Waller & Fawcett, 2013; G. Wang et al., 2016). There are several potential advantages resulting from the adoption of BDA, including an increase of revenues, customer satisfaction and product quality, better resource planning, better insights on customer needs, optimized supply chain, better demand forecast, among others (Lueth et al., 2016). Indeed, previous research (Božič & Dimovski, 2019; Lueth et al., 2016; Trkman et al., 2010) has been highlighting the importance of BDA as a key driver of value creation in modern companies and recent supply chain empirical studies corroborate these findings (Costa & Santos, 2017a; Roßmann et al., 2018; Vieira et al., 2019).

SC is a complex network of entities, processes and resources (Chopra & Meindl, 2016; Harland et al., 2003; Simchi-Levi et al., 2000). Globalization and the global market competitiveness have contributed to making supply chains a more complex network (Harland et al., 2003; Manavalan & Jayakrishna, 2019). Yet, due to this increase of complexity, supply chains have become more vulnerable to disruptions caused by several uncertainties and risks, such as lead-time uncertainty, demand uncertainty, price uncertainty, yield uncertainty, order crossover, supply disruption and supplier delays, which impact on supply chain performance (Er Kara et al., 2020; Fahimnia, Tang, et al., 2015; Harland et al., 2003; C. Tang, 2006b). Over the past two decades, managing these uncertainty factors and risks, and mitigating their effects on supply chain performance have increasingly attracted the attention of researchers and practitioners. Supply Chain Risk Management (SCRM) emerged in the early 2000s and has become an important and attractive field (Baryannis, Dani, & Antoniou, 2019; Sodhi et al., 2012; O. Tang & Musa, 2011). According to studies conducted by G. Wang et al. (2016), Tiwari et al. (2018) and Spanaki et al. (2018), several different techniques have been employed to tackle SCRM-related issues, namely statistics, simulation, optimization and ML (Brintrup et al., 2020). Baryannis, Dani, and Antoniou (2019) classified the techniques used in solutions proposed in the research literature into three main categories: multi-criteria decision analysis techniques, mathematical modeling and optimization and AI techniques. From these, Baryannis, Validi, et al. (2019), Baryannis, Dani, and Antoniou (2019) and Brintrup et al. (2020) argue that the optimization technique has been the most popular (mainly mathematical programming) regarding the majority of SCRM-related studies. Yet, the authors raised attention to the importance of exploring ML approaches, which have received little attention from researchers and practitioners when solving general SC-related problems. In particular, Baryannis, Validi, et al. (2019) have found several gaps in the SCRM literature.

Firstly, there is a lack of studies that employ AI/ML techniques in order to conduct descriptive and predictive risk analyses based on learning mechanisms, in sharp contrast to mathematical programming techniques that do not allow such capabilities. Secondly, notwithstanding BDA has been used in several research fields and applications with considerable impact, empirical research is necessary to further exploit its benefits in the context of SCRM. Thirdly, the vast majority of studies in the literature focus on risk response in detriment of the proactive identification of supply chain risks. At this point, we believe that modeling techniques such as ML and BDA may play an important role on identifying risks in a proactive rather than reactive fashion, thereby enhancing supply chain decision-making processes. Note that the recent works of Brintrup et al. (2020) and Baryannis, Dani, and Antoniou (2019) have already discussed the potential of using Big Data to predict supply chain disruptions. However, scarce attention has been given to BDA applications to real-world supply chain contexts (see, e.g., Er Kara et al. (2020) and Vieira et al. (2019)).

4.2.2 Aims and contributions

Focusing on risk management practices in the upstream supply chain, delivery delays have a direct influence on the overall inventory management performance, potentially leading to inventory stock-outs and damages in the target customer service level. On the other hand, supply orders can eventually arrive earlier than required, potentially entailing an increase in holding costs as a result of additional storage space requirements. As such, on-time delivery is a standard objective of logistics and should be properly estimated so that to avoid shortage of inventory and consequently manufacturing disruptions (Baily et al., 2015). In short, improving on-time delivery performance is crucial for the organizations in order to enhance customer satisfaction and company performance (Niemi et al., 2020).

This paper aims to propose a ML framework for predicting the risk of supply delay by taking advantage of a scalable technological Big Data architecture proposed in N. Silva et al. (2021). The proposed ML framework is supported by a BDW that aimed to store, integrate and provide real data to it. We explore several ML classifiers to predict supplier delays, thereby helping logistics planners in the supply decision-making process. This framework is applied and evaluated in a real-world multinational automotive electronics organization. It is noteworthy that, very often, the performance of ML classifier is determined only in terms of its predictive power, assuming that all potential misclassifications have equal costs. In real-world supply chains, this assumption may not hold and it is necessary to take into account the asymmetric classification costs as an important factor (Klawonn et al., 2011). For instance, depending on the type and characteristics of the products, type of suppliers, distance and transportation modes. Motivated by the lack of research studies considering the impact of misclassifications in supply chain performance, we evaluate the performance of ML classifiers in terms of both predictive performance and misclassification-related costs in order to provide results simultaneously with a high prediction performance and minimum costs of misclassification. Regarding the predictive performance, we consider the Area Under the ROC Curve (AUC) of the Receiver Operating Characteristic (ROC) analysis as the evaluation metric, whereas for the misclassification-related costs we develop a heuristic procedure to calculate the special freight costs and inventory costs associated with the different types of predictive errors. The proposed framework reveals valuable for the organization, helping on improving the efficiency of the inventory management process, enabling to generate cost savings and to ensure an appropriate logistics performance and control. At this point, we believe that our approach may help logistics planners from other related supply chain contexts to leverage the decision-making process to a better performance (acting proactively rather than reactively to supply delays) and, consequently, to increase their flexibility and productivity.

The research methodology adopted throughout this work is the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). The CRISP-DM methodology is a robust and well-proven methodology that provides an overview of the life cycle of data mining projects. It is an iterative methodology in the sense of providing continuous improvements of the artefacts and consists of six distinct phases, as depicted in Figure 2 (Chapman et al., 2000).

The main contributions of this work can be summarized as follows:

- we propose a ML pipeline supported by a technological Big Data architecture in order to predict supply delay risk;
- we compare the performance of six flexible ML classification models (Random Forest, Logistic Regression, Gradient-Boosted Tree, Linear Support Vector Machine, Decision Tree, and Multilayer Perceptron);
- we adopt a realist and robust model Rolling Window (RW) evaluation procedure, with several training and test modeling iterations;
- we address a real-world case study related with a major multinational automotive electronics manufacturer;
- we evaluate the performance of the different flexible ML classifiers in terms of both predictive performance and inventory-related costs induced by a potential model misclassification.

The rest of the paper is organized as follows. Section 4.2.3 provides an overview regarding the literature contributions related to supply delays risk. Section 4.2.4 provides details on the industrial case study that supports our work. In Section 4.2.5, we describe the proposed technological Big Data architecture that serves as foundation for the machine learning pipeline, as well as the methods used to create and deploy the Machine Learning models. All results are presented in the Section 4.2.6. Finally, we conclude the paper in Section 4.2.7.

4.2.3 Related work

BDA and ML techniques have been increasingly considered in the literature to solve inbound logistics problems in supply chains (Baryannis, Dani, & Antoniou, 2019; Er Kara et al., 2020). Table 18 provides

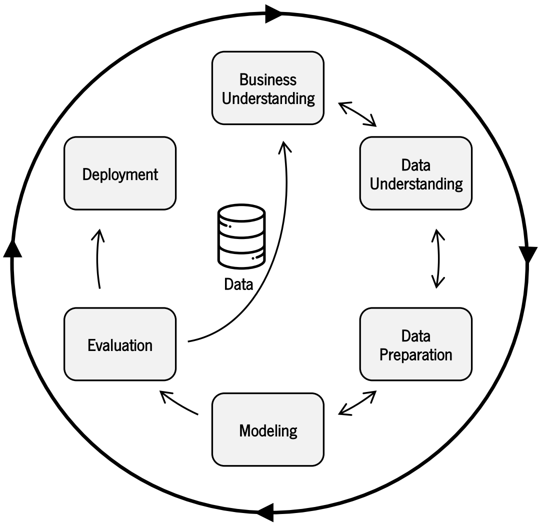


Figure 42: CRISP-DM methodology adapted from Chapman et al. (2000).

an overview of literature contributions on supply delays risk. To the best of our knowledge, there are only six research works in the literature that applied ML techniques for modeling supply delays risk. Quah et al. (2019) proposed a ML model for predicting late deliveries of goods on the Malaysia national courier service. The authors compared three classification models: Naïve Bayes, Decision Tree and K-Nearest Neighbors (K-NN), and used metrics such as precision and recall to evaluated the model predictive performance. Khan et al. (2019) also proposed a Machine Learning model for predicting on-time delivery and late deliveries, comparing also a set of classification models: Support Vector Machines, Logistic Regression, Naïve Bayes and Decision trees. The authors used different performance measures such as accuracy, precision, sensitivity, specificity, F1 score and AUC to evaluate the performance of models. Baryannis, Dani, and Antoniou (2019) proposed and implemented a framework for data-driven risk prediction in a multi-tier aerospace manufacturing supply chain for predicting delivery delays. They applied two ML classification algorithms (Support Vector Machines and Decision Tree) on data related to 50,000 products deliveries from tier 2 to tier 1 supplier in a period of 6 years. The authors also adopted the F1 score, average precision and Matthews correlation coefficient in order to evaluate the models predictive capabilities. The choice of the algorithm was performed considering the trade-off between performance and interpretability of models. Cavalcante et al. (2019) developed a hybrid model that combines simulation and Machine Learning techniques for resilient supplier selection. The authors considered on-time delivery as the key criterion for supplier reliability and compared two classification models: K-NN and Logistic Regression. Recently, Brintrup et al. (2020) used a ML approach to predict on-time and late deliveries on a complex asset manufacturer. The authors compared four ML models (Random Forest, Support Vector Machine, K-NN and Logistic Regression) and empirically evaluated the ML performance using metrics such as precision, recall and F1 score. They also highlight the importance of domain knowledge for performing feature engineering tasks. In a different context, Balster et al. (2020) proposed an Estimated Time of Arrival (ETA) predicting model based on the ML technique for Intermodal Freight Transport Networks (IFTN). This model aimed to predict accurate arrival time on the downstream process in order to increase the supply chain visibility and also to evaluate the impact of delays when they occur. Due to the complexity of this problem, the authors divided the overall ETA prediction into several sub-problems that focus on each lag of the Intermodal freight transport with its appropriate ML model. The overall ETA prediction covers the intermodal transport chain from the origin to the final destination and was obtained by combining each individual prediction.

In sharp contrast with the previous literature, we investigate the performance of the ML-based models by evaluating not only their generalization capacity to unseen data but also the inventory-related costs generated from a potential model misclassification. To the best of our knowledge, this work is the first in the supply chain management literature that combines ML and BD to predict the risk of supply delays while evaluating the financial impact of model misclassifications. In contrast, previous works in the literature only consider ML or the combination of ML with simulation techniques. We intend to fill the aforementioned gap highlighted by both Baryannis, Dani, and Antoniou (2019) and Brintrup et al. (2020) regarding the nonexistence of practical real-world cases that use BD to predict supply chain disruptions (since there are only theoretical frameworks), and the gap pointed by Baryannis, Validi, et al. (2019) related to the lack of studies that employ BDA and ML for identifying risk (instead of focusing on risk response).

4.2.4 Industrial Case Study

This paper aims to propose a Machine Learning approach to predict the risk of supply delays (discussed in Section 4.2.5) by using Big Data technologies (discussed in Section 4.2.5.1) in the Logistics Department of a multinational automotive industry. For the sake of confidentiality, the company name is not disclosed. Currently, the company logistics planners act reactively regarding supply delays, due to the nonexistence of approaches for dealing with this issue. The organization's inbound process managed by the Logistics Department involves raw materials (for instance, different types of components or parts) that will impact in the production of several finished products. The company constantly receives several scheduled deliveries from its suppliers and the production lines run 24h, 7 days per week. Thus, any delay in the delivery of raw materials used in the production plan can lead to delays in the finished goods delivered to customers. These delays will affect the financial results and the organization´s service level.

In order to comply with the finished goods delivery scheduling, if the logistics planners anticipate that a raw material delivery will be delayed, they can act proactively, requesting a special freight that will get the raw material from the supplier to the factory in the necessary time to avoid damages in the production plan. Although the production plan can be fulfilled, the special freight has a substantial monetary cost for the organization. Moreover, if the logistics planner does not anticipate the delivery delay with the necessary time space, the special freight can be late for the planned production. In the other spectrum, the planners can request the special freight due to an assumption of a delay in the supplier delivery but can be a wrong assumption, what will create a monetary cost for the special freight and also for the extra stock. This work addresses a relevant business goal that aims to prevent these situations: to predict

Table 18: Chronological scientific works for modeling supply delays risk using BDA techniques.

Study	**T**[a]	**A**[b]	**M**[c]	**D**[d]	**H**[e]	**E**[f]	**P**[g]	**R**[h]
Quah et al. (2019)	ML	NB, DT, KNN	TPR, PPV	Yes	-	HO	-	400k
Khan et al. (2019)	ML	SVM, LR, NB, DT	ACC, PPV, TPR, TNR, F1, AUC	No	-	CV	-	505k
Baryannis, Dani, and Antoniou (2019)	ML	SVM, DT	F1, CM, MCC, AP	Yes	GS	CV	5y	50k
Cavalcante et al. (2019)	ML, S	KNN, LR	AUC, ACC	No	GS	HO	4y	-
Brintrup et al. (2020)	ML	RF, SVM, LR, KNN	PPV, TPR, F1	Yes	-	HO	1y	232k
Balster et al. (2020)	ML	RF, GBT	RMSE	Yes	GS	CV	2y	131k
This work	ML, BD	DT, RF, LR, GBT, LSVC, MLP	AUC, CM	Yes	BO	RW	20m	60k

[a] Modeling **T**echnique: ML - Machine Learning, S - Simulation, BD - Big Data.

[b] ML **A**lgorithm: NB - Naïve Bayes, DT - Decision Tree, KNN - K-Nearest Neighbour, SVM - Support Vector Machine, LR - Logistic Regression, GBT - Gradient-Boosted Tree, LSVC - Linear Support Vector Machine, MLP - Multilayer Perceptron.

[c] **M**etrics: TPR - True Positive Rate, TNR - True Negative Rate, PPV - Positive Predictive Value, ACC - Accuracy, F1 - F1 Score, AUC - Area Under the ROC Curve, AP - Average Precision, MCC - Matthews Correlation Coefficient; CM - Confusion Matrix.

[d] Empirical **D**ata.

[e] **H**yperparameter Tunning Approach: BO - Bayesian Optimization, GS - Grid Search.

[f] Model **E**valuation: HO - Hold-out, CV - Cross-validation, RW - Rolling Window.

[g] Data **P**eriod: m - month, y - years.

[h] Data **R**ecords: k - thousands of records.

the risk of future planned orders being delayed, supporting the logistics planners decisions to request a special freight in the necessary time frame.

4.2.5 Material and methods

This section discusses the material and methods applied for this research. It includes two main subsections: the first subsection describes the proposed Big Data technological architecture and the second subsection describes the proposed ML pipeline in this research.

4.2.5.1 Big Data technological architecture

Due to the recent advances in IT, there has been a growth of digital data that reflects what is known as the 3 Vs: volume, velocity and variety. Hereupon, the need for new technologies and paradigms arose in order to analyse such data in an adequate time frame. The Big Data paradigm has emerged to solve this issue and, with it, new technologies that use the power of distributed processing and storage. Apache Hadoop is a framework that allows to store data in the HDFS (Shvachko et al., 2010; Vieira et al., 2019;

Zhong et al., 2016) and to process it using the MapReduce (Dean & Ghemawat, 2008; Zhong et al., 2016) programming model. Several of other technologies such as Sqoop, Hive (Thusoo et al., 2009), Spark (Spark, 2018), Impala (Bittorf et al., 2015), among others, are being continuously improved to tackle specific problems in the Big Data ecosystem (Vieira et al., 2019). These technologies allow practitioners to collect data from the data sources, processing and storing it in order to provide useful knowledge to the end-user.

Currently, the amount of Big Data technologies can be overwhelming and sometimes it can be difficult to understand and choose the right technology for the right task. For instance, data collection can be performed using technologies such as Flume, Kafka or Talend. For data preparation and enrichment, we may use Spark, Pig, Storm, or native MapReduce. For data storage, Kudu, Hive, HDFS, or NoSQL databases can be used. For ML tasks, we have Spark, H2O, Petuum, Vowpal Wabbit, Apache SAMOA, and TensorFlow (L'Heureux et al., 2017). Querying data can be performed using Drill, Impala or Presto. For visualization, tools like Tableau, PowerBI or JavaScript can be used (Costa et al., 2018). In this work, due to the analyzed manufacturer requirements and the available technologies in the organization, we propose a technological architecture that is shown in Figure 43. This technological architecture includes three main layers, as follows:

A. **Sources**: refers to the organization's databases or other data repositories. In this work, apart from the transactional SQL database belonging to the organization, we also used one file that was previously stored in HDFS along with additional information related to the suppliers.

B. **Big Data Infrastructure**: based in the Hadoop ecosystem we adopted a set of tools, namely:

   1. **Sqoop** - used to extract data from the organization's SQL databases. This tool only allows the connection to structured data stores (Aravinth et al., 2015). In our case, since the data source is a SQL database there was no need to use another technology for data extraction. After extracting the data from the data source, the same was stored in HDFS using the *Parquet* format. This format is one of several formats available in the Hadoop ecosystem, such as ORC, AVRO, among others (Ivanov & Pergolesi, 2020). We adopted the *Parquet* format due to its compatibility with Apache Spark and Impala, and also because it is a read-oriented format with adequate compression, which will bring advantages when querying the data (Baranowski et al., 2015).

   2. **Apache Spark** - framework for data cleansing, transformation and ML models creation. It provides several libraries, such as *spark.sql*, *spark.mllib* and *spark.ml* (Armbrust et al., 2015; Meng et al., 2016). The *spark.sql* library allows the use of SQL functions together with the Spark functional programming API. The *spark.mllib* and *spark.ml* libraries are used for ML purposes. Being able to perform all these tasks in a single framework is an enormous advantage, in the sense that there is no need to spend time configuring and adjusting different

technologies. Moreover, Apache Spark is compatible with the *Parquet* format and Hive, which will be used to provide the data to the end-users.

3. **Hive** - includes the Hive Metastore (the system catalog) where the metadata (schema and other statistics) are stored, allowing query optimization and data exploration (Thusoo et al., 2009). Hive allows the creation of external tables where data is stored in HDFS directories, while the schema and other information are stored in the Hive Metastore (Thusoo et al., 2009; Vieira et al., 2019). Due to this fact, data can be queried using *spark.sql*, Impala, among other technologies.

4. **Impala** - provides a Massively Parallel Processing (MPP) SQL engine that combines the flexibility and scalability of Hadoop with the familiarity of SQL (Bittorf et al., 2015). Furthermore, Impala can query data from Hive and can provide a connection to visualization technologies, such as Tableau or Power BI.

C. **Data visualization**: consists in data visualization tools for analysing the massive amount of data. We adopted Power BI to develop dashboards for the end-users.

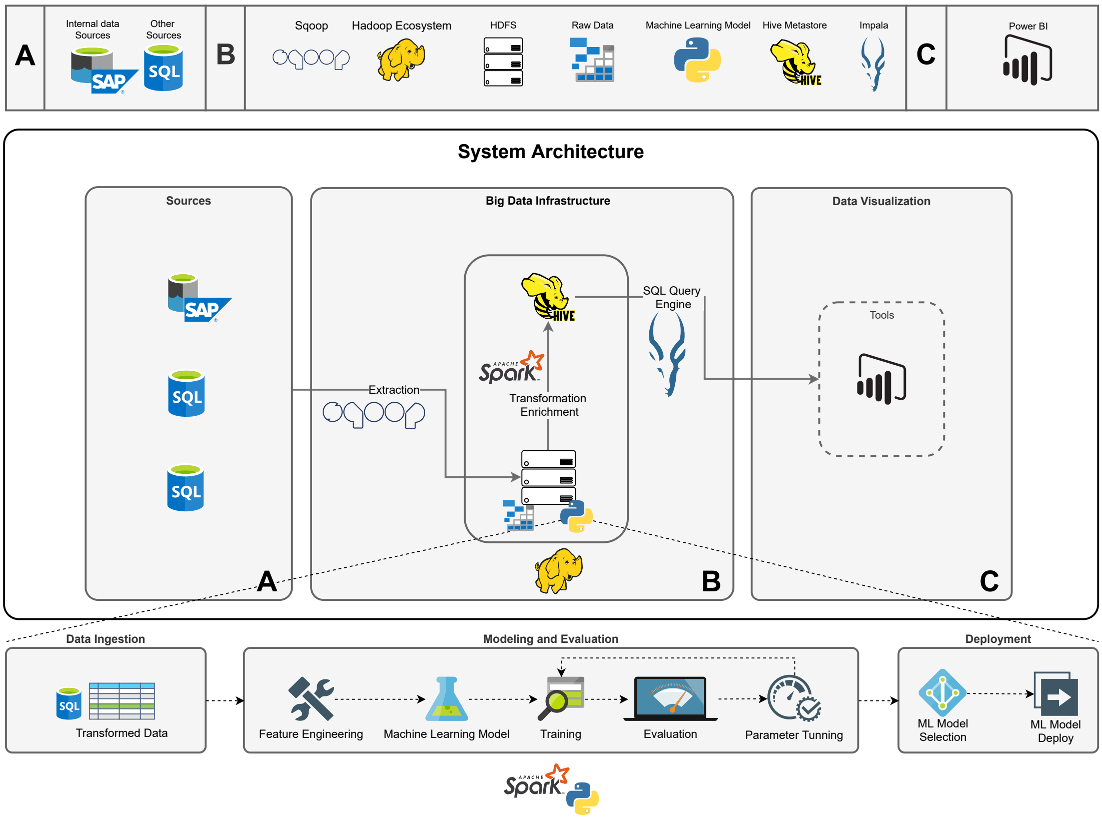


Figure 43: Big Data technological architecture.

We adopted Apache Spark due to its scalability, simplicity and easy integration with other tools. It is a widely-used unified analytics engine for large-scale data processing used into 80% of the Fortune 500 enterprises. It provides fast distributed computing on large-scale data across clusters of machines using in-memory processing. Moreover, Apache Spark supports several high-level tools, as such Spark SQL (for SQL and structured data processing) and MLlib (for building Machine Learning Pipelines in large-scale environment) (Spark, 2018). We also adopted the Sqoop tool for the data extraction data due to the fact that all sources are SQL databases (relational databases). The choice for Impala as the SQL query engine was based on its performance on query execution. Generally, Impala assuring low latency query execution compared with other SQL-on-Hadoop systems, such as Spark, Drill, Presto or Hive (M. Santos et al., 2017; N. Silva et al., 2021). Finally, we choose the PowerBI tool for data visualization. It is a powerful, popular and easy to use (drag-and-drop functionality to create visualizations and intuitive features) data visualization tool. Moreover, this tool is recently listed as a Leader in the *2021 Gartner Magic Quadrant for Analytics and Business Intelligence Platforms* (Richardson et al., 2021).

The developed technological architecture fulfills the requirements of this project ensuring that the knowledge provided by the data analysis can be used by the end-users to support their decision-making processes and to improve the organization's results.

4.2.5.2 Machine Learning pipeline

The proposed ML pipeline is shown in Figure 44 and details all steps followed to developed the final models. It includes three main steps:

- **Data Preparation:** consists in activities such as data extraction from data sources and perform all tasks related to data cleaning, data transformation (including, feature engineering), as well as standardization of features and feature selection. This step is outlined in Section 4.2.5.3;

- **Modeling and Evaluation:** the ML model is built and then it predictive capability (performance evaluation) is evaluated. The detailed specification of this step is provided in Sections 4.2.5.4, 4.2.5.5 and 4.2.5.6; and

- **Deployment:** the selected ML model is deployed on Big Data cluster. Firstly, the model is saved into a specific path of the HDFS and then is created an workflow to re-train this model in a predefined time-frame. All predictions are stored in the Hive external tables;

4.2.5.3 Data Preparation

4.2.5.3.1 Logistics data

This research considers real data from the Logistics Department of an automotive industry, from January 2019 to September 2020, related to orders of raw material. The data were collected from the company data repository (system that contains the same information of the *ERP* system) using the *Sqoop* technology, and then stored in the Big Data cluster in a *Parquet* format. Table 19 provides a description of

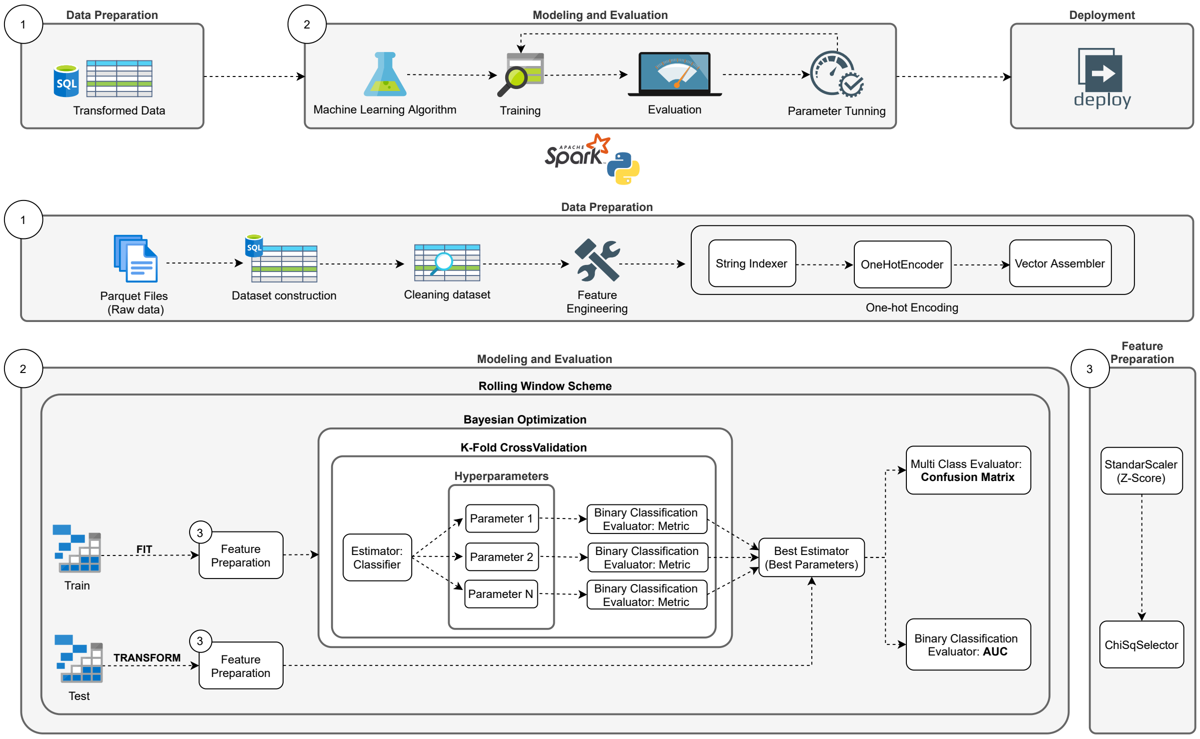


Figure 44: Machine learning pipeline.

the constructed dataset, as well as the identification of all features selected as the input of the predictive model. These features were selected manually in concordance with business domain expert knowledge, as defined by the logistic planners. In addition, a set of new features were created from the raw data to improve the predictive capabilities of the ML models (see Section 4.2.5.3).

After collecting data, we perform the data cleansing in order to remove all missing and duplicated data. Therefore, we obtained a total of 60,340 orders related to 3,414 different raw materials, from 259 suppliers. Performing the Exploratory Data Analysis (EDA), we found that 60% of orders were related to deliveries with delays (class “0”) and 40% of orders (class “1”) were related to early and on-time deliveries. Thus, our dataset is reasonably balanced, as shown in the Figure 45. We defined that the negative class (“0”) corresponds to order delays and the positive class (“1”) corresponds to a early and on-time delivery orders.

Several ML algorithms require only numerical features and therefore categorical features must be previously converted into numerical features. In this work, five of the considered features are categorical and we applied the known one-hot encoding to convert them to numerical values. Firstly, we applied the *StringIndexer* to encodes each column of labels into column of label indices. Then, the label indexes are mapped to a binary vector using the *OneHotEncoder*. Lastly, we use the *VectorAssembler* to merge these vectors into a single feature vector.

Before encoding categorical features we must verify and visualize each categorical variable and its levels, in order to determine whether or not it should be considered as a feature for our models and subsequently encoding them if necessary. We noted that all of our categorical features levels occur in a

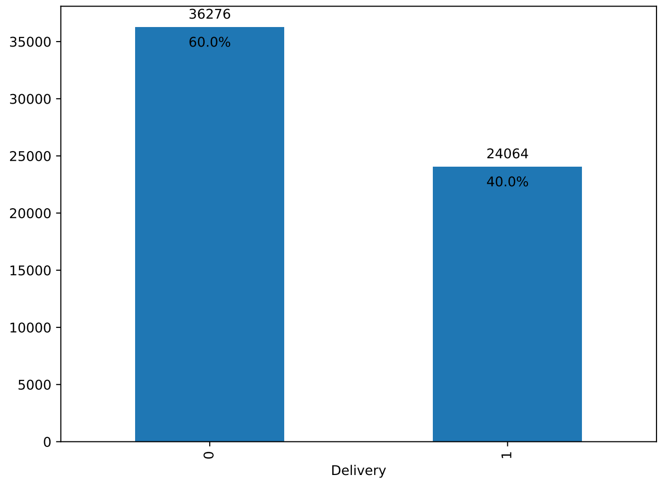


Figure 45: Distribution of output target.

considerable proportion over the dataset (variations in the dataset), making some impact on the model.

Table 19: Summary of the logistics data attributes.

Context	**Attribute**	**Description (format, examples)**	**F**[a]
Orders	Order date	Order placed date (date, {'20190320'})	N
	Order quantity	Quantity of ordered item (number, {1800, 10000})	Y
	Quantity delivered	Quantity delivered (number, {1500, 8000})	N
	Delivery date (Planned)	Planned date to deliver the order (date, {'20190802'})	N
	Delivery date (Real)	Date of order delivery (date, {'20200302'})	N
	Transport mode	Type of transportation (text, {'Sea', 'Air', 'Road'})	Y
	Forwarder	Name of delivering carrier (text, {'Company X'})	N
Raw materials	Material number	Raw material code (text, {'00126523', 'FH123201'})	N
	Lead-time	Contracted transit time in working days (number, {1, 4})	Y
	Planning time fence	Frozen zone in days at supplier (number, {2, 8})	Y
	ABC classification	Classification of raw material determined by volume of sales and price (text, {'A', 'B', 'C'})	Y
Suppliers	Number	Code of supplier (text, {'0685RF5T'})	N
	Name	Legal name of the supplier (text, {'Company Y'})	N
	Location (Country)	Location country of supplier (text, {'Portugal', 'China'})	N
	Location (Region)	Region of the supplier location (text, {'Europe'})	Y
Target	Score	Supplier score: delays and on-time delivery (binary, {0, 1})	Y

[a] **F**eature: Y - Yes, N - No.

4.2.5.3.2 Feature engineering

Feature engineering is a fundamental process in any ML task. Often, it involves the creation of a new set of features (or attributes) from raw data that enable a more effective characterization of the predictive problem (Domingos, 2012). Inspired by the work of Brintrup et al. (2020), and after several brainstorming meetings with the company logistics experts, we have built a set of new features from the raw data (considered as inputs in the predictive modeling process) as follows:

- **Percentage of delays**: Naturally, we expect that the past on-time-delivery performance of the supplier may influence future deliveries. Therefore, we have considered the average percentage of supplier delays over the past n orders, i.e.,

$$Delays(t)^{(n)} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{A_{t-i} > R_{t-i}\}} \times 100 \tag{4.1}$$

  where A_{t-i} and R_{t-i} are the actual and required receipt dates for the past supplier order $t-i$, and 1_A denotes a logical function (it equals 1 if A is true and 0 otherwise). Note that t corresponds to the index of the supplier order under prediction. The structure of Eq. (4.1) implies that the first n instances of the data are dropped due to the use of time lags of order n.

- **Average past/future order volume**: The level of volatility of manufacturer's orders to the suppliers may impact on their flexibility to cope with the manufacturer's needs. Hence, we consider the average quantity ordered over the past n orders as input feature for our models, i.e.,

$$AVGPastVol(t)^{(n)} = \frac{1}{n} \sum_{i=1}^{n} O_{t-i} \tag{4.2}$$

  This feature can be an indicator of the pressure exerted on the supplier over the last recent orders. Likewise, we have also considered the average future order volume that remains undelivered and may impact on the current supplier's ability to deliver (or not) on the targeted date, i.e.,

$$AVGFutureVol(t)^{(n)} = \frac{1}{n} \sum_{i=1}^{n} O_{t+i} \tag{4.3}$$

- **Supplier flexibility**: The variation, in volume, of subsequently orders (Eq. 4.4) may be a relevant indicator regarding the supplier's flexibility to cope with short-term order quantity variations.

$$SuppFlex(t) = O_t - O_{t-1} \tag{4.4}$$

In other words, we expect that there is a relationship between short-term large variations in supplier orders and supply delays. Following this reasoning, to account for mid-term variations on the order quantities, rather than just short-term dynamics, we further extend the expression (4.4) over the past n orders as follows:

$$OrderDelta(t)^{(n)} = O_t - \frac{1}{n}\sum_{i=1}^{n} O_{t-i} \tag{4.5}$$

In this sense, we can explicitly account for events where order variations are high in the short-term but smoothed in the mid-term, and vice-versa. Currently, the case-study company takes advantage of a score metric to evaluate the performance of the supplier. This metric, hereinafter called *score*, can be understood as an indicator function that takes the numerical value 1 if the order arrives on-time with the planned quantity and 0 otherwise. Apart from this input feature, we also consider the average supplier score over the past n orders in our modeling experiments, i.e.,

$$AVGScore(t)^{(n)} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{A_{t-i}=R_{t-i}\}} \tag{4.6}$$

In this work, we set the maximum time lag order to $n = 10$, as it is considered to be a sufficiently long time frame to realistically evaluate the supplier's performance by the company experts. Notwithstanding, it should be emphasized that this value can naturally vary depending on the case study under consideration.

Apart from the above features, we also consider some core master logistics information related to the raw material to be ordered (see Table 19), including its: (i) ABC classification, (ii) production and shipping region, (iii) transportation mode and (contracted) lead time and (iv) frozen zone length (period in which no changes to the planned orders are allowed). In addition, seasonality information implicitly embedded in the order's month and season of the year is also incorporated as potential features to our classifiers.

4.2.5.3.3 Feature selection

Feature selection is an core task in data mining projects, providing several benefits such as reducing the number of features and storage requirements, reducing both training and utilization times, reducing the probability of overfitting, improving the prediction performance, among others (Domingos, 2012; Guyon & Elisseeff, 2003). It consists of selecting a subset of features to be used on the predictive model, removing all irrelevant and redundant features that cause a negative effect on several Machine Learning schemes (Witten et al., 2016). According to Guyon and Elisseeff (2003), there are three different methods for features/variables selection: wrappers, filters and embedded. Wrappers methods define the subset of features according to their predictive power in the selected Machine Learning algorithm. In filters methods, the subset of features is selected during the preprocessing stage without predefined a learning algorithm. Lastly, embedded methods define the subset of features during the training process with a predefined learning algorithm.

Witten et al. (2016) argues that the manual selection based on the problem domain knowledge is the best method for selecting relevant attributes, warning also to the usefulness of automatic methods. In this work, we adopt both strategies for feature selection. Firstly, we perform the manual selection using the knowledge of the business domain expert (logistic planners), resulting in a set of 27 selected features. After this first stage, we define the *ChiSqSelector* from *spark.ml* library as our automatic selection approach (filter method). *ChiSqSelector* is based on the Chi-Square test of independence and it is useful when dealing with categorical features, measuring the independence between them over the class labels of the target variable. This feature selector supports the following five selection methods (Spark, 2020):

- *numTopFeatures* - selects a defined number of top features based on the chi-squared test;
- *percentile* - selects a percentage of top features based on the chi-squared test;
- *fpr* - selects all features with a p-value below the defined threshold, in order to control the false-positive rate of selection;
- *fdr* - selects all features with a false discovery rate below the defined threshold (uses the *Benjamini-Hochberg* procedure);
- *fwe* - selects all features with a p-value below the defined threshold (1/numFeatures), in order to control the family-wise error rate of selection.

In this work, we perform a preliminary experiment to determine the number of features to be selected. First, we define a set of candidate number of features and then each one is tested through the *numTopFeatures* parameter of *ChiSqSelector* method. Afterwards, the number of selected features is determined from the predictive performance of models in terms of the AUC metric (Fawcett, 2006).

4.2.5.4 Classification algorithms

We tested six ML classification algorithms: Decision Tree (DT), RF, Logistic Regression (LR), Gradient-Boosted Tree (GBT), Linear Support Vector Machine (LSVC) and Multilayer Perceptron (MLP). Following the *spark.ml* library, we have considered a set of relevant hyperparameters for each classification model, as described in Table 20.

4.2.5.4.1 Decision Tree (DT)

A decision tree consists of a collection of decision nodes connected by edges or branches that extends downwards from the root node until the terminal or leaf nodes (Larose, 2005; S. Li & Zhang, 2020). It aims to classify instances, sorting them based on feature values (Kotsiantis et al., 2006). The attributes are tested at decision nodes, where each possible value results in a branch. Each branch can lead to another decision node or to a terminating leaf node and each leaf node represents class labels associated with the instance (Kotsiantis et al., 2006; Larose, 2005; S. Li & Zhang, 2020). Instances are classified by starting from the root node extending downwards to a leaf, according to the outcome of the tests along the path (Kotsiantis et al., 2006; Luo, 2022).

4.2.5.4.2 Random Forest (RF)

Random Forest is a popular classification and regression method developed by Breiman (2011). It combines bagging technique also known as bootstrap aggregation, with random feature selection to build assembles of decision trees (multiple models of several decision trees) in order to reduce the risk of overfitting and achieve better prediction performance (Spark, 2020; Witten et al., 2016).

4.2.5.4.3 Logistic Regression (LR)

Logistic regression is a popular ML algorithm widely applied to predict a binary outcome. Logistic regression is defined as:

$$p(C = 1|x) = \frac{1}{1 + exp(-(\beta_0 \sum_{i=1}^{n} \beta_i, x_i))} \tag{4.7}$$

where, x is the instance to be classified, and β_0, β_1, ..., β_n represent the parameters of the model. These parameters should be estimated from the data (Hosmer et al., 2013). The *spark.ml* library provides two algorithms for solving logistic regression: mini-batch gradient descent and Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS).

4.2.5.4.4 Gradient-Boosted Tree (GBT)

The Gradient-Boosted Tree (GBT) is a classification and regression method, which uses the boosting technique to build assemblies of decision trees. The boosting method aims at converting weak learners into strong learners in a iterative fashion. It applies weak learners sequentially, in order to repeatedly re-weight versions of the training data (Hastie et al., 2009; Krauss et al., 2017).

4.2.5.4.5 Linear Support Vector Machine (LSVC)

Support Vector Machine (SVM) is a supervised Machine Learning algorithm used for classification and regression purposes (Burges, 1998). It use the nonlinear mapping to transform the input $x \in \mathbb{R}^m$ into a high *m*-dimensional feature space with kernel function. Then, it finds the optimal linear separating hyperplane according to a set of support vector points in the feature space (Cortez, 2010). The optimal separating hyperplane corresponds to the hyperplane with the largest distance to the nearest training data points of any class, known as functional margin. In general, the larger is the margin, the lower is the generalization error of the classifier. For the LSVC, *spark.ml* package optimizes the Hing Loss using Orthant-Wise Limited-memory Quasi-Newton (OWL-QN) optimization algorithm (Spark, 2020).

4.2.5.4.6 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a feedforward artificial neural network based classifier that consists of at least three layers: one input layer, one or more hidden layers, and one output layer. The number of input layer nodes (or neurons) is defined by the input data and the output layer by the number of classes

(Larose, 2005; Luo, 2022; Spark, 2020). The number of hidden layers, as well as the number of nodes in each hidden layer, are both defined by the user based on the particular case on hand (Dwivedi, 2018; Larose, 2005). An multilayer perceptron network without hidden layer is equivalent to multinomial logistic regression (Hastie et al., 2009). The hidden layer nodes considered the sigmoid (logistic) function (Spark, 2020):

$$f(zi) = \frac{1}{(1 + exp(-zi))} \tag{4.8}$$

and nodes in the output layers considered the softmax function (Spark, 2020):

$$f(zi) = \frac{exp(zi)}{\sum_{k=1}^{N} exp(zk))} \tag{4.9}$$

In this work, we adopt a MLP network with one input layer composed by x neurons, one hidden layer with H neurons and two neurons (binary classification) in the output layer. For MLP, spark.ml library optimizes the the loss function using the L-BFGS algorithm. The L-BFGS is a quasi-Newton method that conducts an approximation of the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm but adapted to use a limited amount of computer memory.

Table 20: Hyperparameter of the classification algorithms

Parameter	**Classification algorithms**						**Description**
	DT	**RF**	**LR**	**GBT**	**LSVC**	**MLP**	
maxDepth	✓	✓		✓			Maximum depth of the tree (*d=5*)
maxBins	✓	✓		✓			Maximum of bins for discretizing continuous features (*d=32*)
numTrees		✓					Number of trees to train (*d=20*)
elasticNetParam			✓				ElasticNet mixing parameter in a range of [0, 1](*d=0.0*)
regParam			✓		✓		Regularization parameter (*d=0.0*)
maxIter			✓	✓	✓	✓	Maximum number of interactions (*d=100*)
layers						✓	Size of layers: input, hidden and output layers (*d=None*)

[a] Default Value (d)

4.2.5.5 Hyperparameter Tuning

Hyperparameter tunning is a key component for building effective ML models. In this work, we adopt a Bayesian optimization to search for the best ML hyperparameters by using the *HyperOpt* library[1]. Unlike grid search and random search, Bayesian optimization can detect the optimal hyperparameter combination within fewer interactions and also determine the next hyperparameter based on the previously-evaluated outcomes, reducing the unnecessary evaluations and also improving the efficiency (L. Yang & Shami, 2020).

[1] http://hyperopt.github.io/hyperopt/

Apache Spark provides two built-in tools for hyperparameter optimization: Cross-validation and Train-validationSplit. Both uses grid search (parameter grid) for determining the optimal parameters for a given model. However, we integrate *HyperOpt* with Apache Spark *CrossValidator (k-fold cross-validation)* for hyperparameter tuning. The *HyperOpt* library is an open-source python library created by James Bergstra based on Bayesian optimization and provides three tuning algorithms (Bergstra et al., 2011): Tree of Parzen Estimators (TPE), Adaptive Tree of Parzen Estimators (ATPE) and Random Search (RS). Moreover, HyperOpt requires four conceptual components (Bergstra et al., 2013):

- **Function**: the objective function to be minimized that receives hyperparameters from the search space and return the *loss* (criterion to be minimized);
- **Search space**: the parameter space for searching;
- **Algorithm**: the optimization algorithm; and
- **Evaluations**: the maximum number of evaluations.

In this work, first we define the hyperparameter space for each models tested (Section 4.2.5.4). Then, in the objective function we employed a k-fold cross-validation to further define the AUC metric as the criterion to be minimized or *loss* (symmetrical of the AUC metric: $-1 \times AUC$).

4.2.5.6 Evaluation

We use a realistic and robust RW scheme to evaluate the classification models as illustrated in Figure 46. This scheme is realistic in the sense that simulates the real environment in which a model would be used, producing several training and test iterations over time. And, is it robust because each iterations produces a set of predictions, thus there are several model evaluations through time, as opposed to the popular single hold-out train and test scheme.

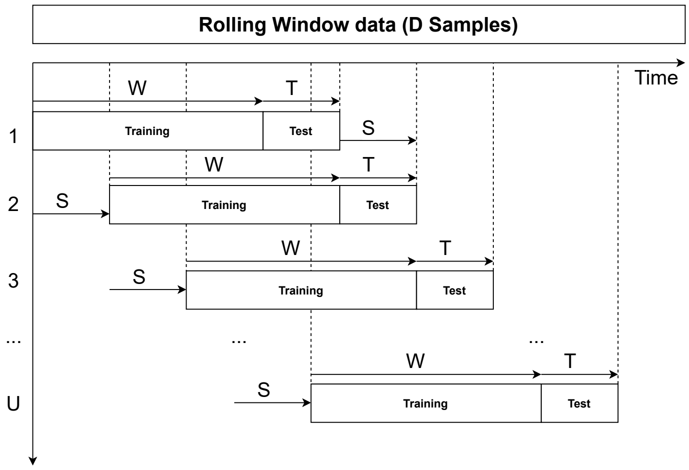


Figure 46: Rolling window scheme adapted from (Oliveira et al., 2017)

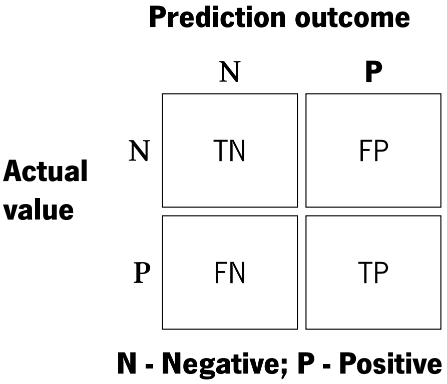
This scheme works as follows. In the first iteration ($U = 1$), the model is trained using a fixed training window W with the oldest samples and predicts the subsequent T samples. Then, in the second iteration ($U = 2$), the training window slides in S instances, causing the replacement of the S oldest instances of the training window by the S recent ones. A new model is fit and then predicts the new subsequent T samples. This process is repeated until the last interaction of the RW scheme. The total of iterations is determined using the following formula:

$$U = (D - (W + T))/S.$$

4.2.5.6.1 Measuring model performance

The evaluation criteria are a key factor for quantifying the model performance (Galar et al., 2012; Witten et al., 2016). In this work, the overall performance of classification models is given by the AUC of ROC analysis, also known as AUC or AUC-ROC (Fawcett, 2006). The ROC analysis is obtained by considering the predictions as probabilities (p) for a binary class. The class is assumed true if $p > D$, where D is a decision threshold. When using a fixed D, the predicted class labels can be used to compute the well known the confusion matrix. Figure 47 illustrates an example of such matrix, which matches predicted outcomes with the actual values and includes four main statistics for the binary classification task (Spark, 2020):

- True Positives (TP) - number of positive class correctly classified;
- True Negatives (TN) - number of negative class correctly classified;
- False Positives (FP) - number of negative class incorrectly classified as positive class and;
- False Negatives (FN) - number of positive classes incorrectly classified as negative class.



Prediction outcome

		N	P
Actual value	N	TN	FP
	P	FN	TP

N - Negative; P - Positive

Figure 47: Confusion matrix for a binary classification task.

Several statistics and insights can be obtained from the confusion matrix. Metrics such as True Positive Rate (TPR), True Negative Rate (TNR), Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), F1 score and accuracy are defined using the following formulas (Larose, 2005; Sun et al., 2009):

- $TPR = \frac{TP}{TP+TN}$, also known as recall, hit rate or sensitivity;
- $TNR = \frac{TN}{TN+FP}$, also known as specificity or selectivity;
- $PPV = \frac{TP}{TP+FP}$, also known as precision;
- $FPR = \frac{FP}{FP+TN}$, also known as fall-out;
- $NPV = \frac{TN}{TN+FN}$;
- $ACC = \frac{TP+TN}{TP+TN+FP+FN}$;
- $F1 = 2 * \frac{PPV*TPR}{PPV+TPR}$.

The ROC curve is a two-dimensional graphical representation technique for visualizing, organizing and selecting classifiers based on their performance. It is a curve that summarizes the trade-off between the TPR (*y*-axis) and FPR (*x*-axis) for different threshold points (*D*) between 0.0 and 1.0 (Fawcett, 2006; Sun et al., 2009). The AUC-ROC measures the quality of the probabilistic classifier and is calculated using Equation (4.10). A random classifier has AUC-ROC of 0.5, while a perfect classifier has AUC-ROC of 1.

$$AUC - ROC = \int_0^1 \frac{TP}{TP+FN} d \frac{FP}{FP+TN} d = \int_0^1 \frac{TP}{P} d\left(\frac{FP}{N}\right) \tag{4.10}$$

4.2.5.6.2 Measuring misclassification impact on inventory performance

In order to measure the impact of a model misclassification on the inventory performance of the concerned company, we design a cost matrix that determines the cost of classifying samples from one class as another, as shown in Table 48. Following the notation of Elkan (2001) and Sun et al. (2009), $C(i, j)$ denotes the cost of predicting an instance from class *i* as class *j*. Hence, $C(1, 0)$ represents the cost of misclassifying a positive instance as a negative one, whereas $C(0, 1)$ is the cost of misclassifying a negative instance as a positive one.

In this work, the cost matrix is based on inventory-related costs, namely the special freight costs calculated using the business domain expert's knowledge and unitary holding costs. Inventory Holding Costs (IHC) are costs incurred to hold inventory and include capital costs and storage costs. It is calculated using the following formula:

$$IHC_m = (I_m \times P_m) \times V_m \tag{4.11}$$

Prediction outcome

		N	P
Actual value	N	C(0,0)	C(1,0)
	P	C(0,1)	C(1,1)

N - Negative; P - Positive

Figure 48: Cost matrix for the binary classification task.

where I_m is the holding rate for raw material m per unit of time, P_m denotes the raw material standard unit price and V_m represents the order volume. On the other hand, special or premium freight is a type of shipment offered by transportation providers for urgent deliveries. This type of shipments tends to be very expensive and normally performed by airways. In general, the special freights are caused by inventory mismanagement that leads to high stockout risk (Avci & Selim, 2017).

In the context of the case study company, the calculation of the special freight costs follows specific business-oriented rules essentially dependent on two factors: the supplier location and the transport mode. In addition, the company sets a penalty cost that differs according to each combination of these factors. Concretely, national special freights are typically made by land, where the carriers define the price from the number of load units to be transported. Here, the penalty cost ($s_{n,l}$) depends on the distance in kilometers from the company to the national supplier. For in-Europe special freights, the carriers provide shipments by land and air. For shipments by land, the price is defined by the distance from the company. In contrast, for shipments by air, the weight to be transported is the main cost driver. In both cases, the magnitude of the penalty costs $s_{i,l}$ and $s_{i,a}$ is dependent on the urgency required for receiving the supply order. Finally, Out-Europe special freights, typically more costly to manage, are made by air and the corresponding cost function is also determined from the weight to be transported and the transit time necessary for shipping the order. Table 21 summarizes the algebraic expressions used by the analyzed company to determine the special freight costs.

Recalling the cost matrix presented in Fig. 48, it is noteworthy that while the costs related to $C(1,0)$ involve only the special freight component, those related to $C(0,1)$ comprise both the special freight cost as well as the holding cost component. The latter case is motivated by the fact that the classifier triggers a need to carry out a special freight that is unnecessary in light of the actual production requirements. As such, in addition to the special freight cost, the corresponding chartered quantity will cause an increase in the inventory on-hand, which represents an extra stock to be stored.

Table 21: Special freight cost functions according to the transportation mode and supplier location.

Supplier location	Transport mode	Cost function
National	Land	$C = N_q \times P_p \times s_{n,l}$
In-Europe	Land	$C = D \times P_{km} \times s_{i,l}$
	Air	$C = W_q \times P^E_{kg} \times s_{i,a}$
Out-Europe	Air	$C = W_q \times P^O_{kg} \times s_{o,a}$

Legend: C - special freight cost; q - quantity to be transported; N_q - number of load units required to transport q; W_q - total weight (packaging weight + loading weight) of quantity q; D - distance from the company to supplier; P_p - price per load unit; P_{km} - price per kilometer; P^E_{kg} - price per kilogram for suppliers located in Europe; P^O_{kg} - price per kilogram for suppliers located out of Europe; $s_{n,l}$ - penalty cost for national express services carried out by land; $s_{i,l}$ - penalty cost for In-Europe express services carried out by land; $s_{i,a}$ - penalty cost for In-Europe express services carried out by air; $s_{o,a}$ - penalty cost for Out-Europe express services carried out by air.

4.2.6 Experiments and Results

This section describes the experimental and modeling setup, as well as the results of model performance comparison and misclassification cost.

4.2.6.1 Experimental and modeling setup

The experimental setup defined to conduct the experiments is described in Table 22. Regarding the classifiers tested, we explored six different ML algorithms (Section 4.2.5.4) using the Rolling Window evaluation scheme. After consulting domain experts, we determined the values *W = 42.000*, *T = 857* and *S = 857* for the RW scheme, producing a total of *U = 20* iterations. We used the first iteration (*U = 1*) to perform preliminary experiments for the feature selection and also for hyperparameter tuning. Before fitting the ML models, all features were standardized using the standard score, also known as Z-Score standardization. The mean and the standard deviation were calculated over each of the features present in the training set only. For the feature selection, we adopt the hold-out scheme using only the training data. Under this holdout scheme, the training data is divided using a random split of 80% for the training set and 20% for the validation set. Using the *ChiSqSelector* method of the *spark.ml* library, we determined a set of candidate features numbers through the *numTopFeatures* parameter as follows: *numTopFeatures* $= \{27, 24, 20, 18, 15, 13, 8, 4\}$. The results of preliminary experiments are summarized in Table 24. The results show that there is no significative benefit in applying a feature selection, since the computational effort does not substantially reduced, while the predictive performance decreases (AUC values). Thus, we do not use the automatic feature selection and adopt all 27 input features in the remainder ML experiments.

After the feature selection experiments, we compared the performance of six explored Machine Learning models: DT, RF, LR, GBT, LSVC and MLP. We integrated the Bayesian Optimization (*HyperOpt* library) with a k-fold cross-validation (*spark.ml* CrossValidator) for the hyperparameter tuning. The hyperparameters have to be tuned in order to select a set of optimal hyperparameters for a learning algorithm. For each model, we defined the hyperparameter space, as well as the objective function to be minimized over 10 iterations (maximum number of evaluations) of the Tree of Parzen Estimators method. The objective function employed a 5-fold cross-validation to further calculate the AUC metric, which was defined as the criterion to be minimized in the objective function. The DT requires two parameters to be optimized: *maxDepth* and *maxBins*. We defined a hyperparameter space under the ranges *maxDepth* $\in \{2, 5, 10, 20, 30\}$ and *maxBins* $\in \{10, 20, 40, 80, 100\}$. In the case of RF, we set the number of trees to train *numTrees = 200* and we defined the same hyperparameter space of DT, due to it requires the same parameters to be optimized. The GBT was trained with 100 epochs (maxIter = 100) of L-BFGS algorithm and also requires the same parameters of DT to be optimized. Equally, the LR was trained with 100 epochs (maxIter = 100). It required two hyperparameters to be optimized: *elasticNetParam* and *regParam*. We defined a hyperparameter space under the ranges *elasticNetParam* $\in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ and *regParam* $\in \{0.01, 0.1, 0.5, 1.0, 2.0\}$. Lastly, for the MLP, we defined a network of one input layer with 27 neurons (number of input features), one hidden layer with *H* neurons and one output layer with 2 neurons, where *H* is defined using the heuristic $H = round(N/2)$, where *N* is the number of inputs. After setting the best hyperparameters, we retrain the models in order to obtain the final model performance. For each iteration of RW, we standardized all data features using the mean and the standard deviation calculated in the first iteration and we also fixed the best hyperparameter obtained from the hyperparameter tunning.

Lastly, we store the AUC produced in each iteration of the RW, as well as the computational effort (in seconds). Afterwards, we aggregate the distinct AUC values (one for each RW iteration) to calculate the median value for each model. Moreover, we adopted the Wilcoxon non-parametric statistic in order to verify if paired median differences are statistically significant.

4.2.6.2 Overall predictive performance

The overall predictive performance for the six models tested is summarized in Table 25. Figure 49 plots the ROC curve for the tested models in each iteration of RW scheme. The results show that the RF model provides the best predictive capacities compared with other tested models, with a median AUC score of 90%. The GBT provide also good predictive performance with median AUC of 82%, however, it is the slowest model in terms of the training phase (requiring more computational effort). Regarding

Table 22: Experimental setup.

Setup	Description	Specification[a]
Hardware: Cluster	Nodes Cores Disk capacity Memory capacity	9 152 (144 with Hyper-threading) 224 TB 1603 GB
Implementation	Framework	Apache Spark 2.4.0
	Language	PySpark (Python API): The Apache Spark provides high-level API in Java, Python and R programming languages (Spark, 2020).
	Extra packages	*HandySpark*: used to extends evaluation metrics for binary classification (e.g., compute metrics over threshold values (D), among others). *HyperOpt*: used for hyperparameter tuning (Bayesian Optimization).
Execution	Submitting application	spark-submit: used to run all spark applications on the cluster. The adopted configurations are described in Table 23.

[a] Nomenclature: TB - terabyte, GB - gigabyte, API - Application Programming Interface.

Table 23: Configuration of the *spark-submit* parameters (Spark, 2020).

Parameter	Description	C[a]
master	The cluster manager: Standalone, Apache Mesos, Hadoop Yarn or Kubernetes.	yarn
deploy-mode	The deploy mode to run the driver process: *cluster* mode (inside of the cluster) or *client* mode (outside of the cluster, i.e., locally).	cluster
executor-cores	The number of cores to be used for each executor.	4
num-executors	The number of executors.	7
executor-memory	The amount of memory to be used for each executor process.	10GB
driver-memory	The amount of memory to be used for the driver process.	10GB

[a] **C**onfiguration: GB - gigabyte.

Table 24: Results of feature selection experiments in terms of AUC.

Models		**numTopFeatures**							
		27	24	20	18	15	13	8	4
RF	AUC	0.876	0.875	0.871	0.868	0.714	0.684	0.587	0.571
	ET[a]	28.49	26.10	25.99	25.21	25.16	24.98	25.35	25.19
LR	AUC	0.861	0.861	0.860	0.858	0.695	0.683	0.582	0.567
	ET[a]	28.78	26.37	26.05	25.72	25.43	25.31	25.22	24.47
GBT	AUC	0.892	0.892	0.885	0.881	0.740	0.686	0.587	0.571
	ET[a]	29.68	28.64	26.75	25.82	23.75	23.46	22.55	21.56
LSVC	AUC	0.838	0.829	0.833	0.832	0.693	0.671	0.582	0.559
	ET[a]	21.03	20.14	20.80	20.16	20.64	28.97	22.67	20.98
MLP	AUC	0.877	0.877	0.874	0.871	0.716	0.688	0.587	0.571
	ET[a]	23.45	21.31	20.64	20.45	19.35	18.90	18.26	15.42
DT	AUC	0.857	0.857	0.845	0.855	0.706	0.681	0.582	0.569
	ET[a]	26.24	25.43	25.29	25.58	24.33	24.74	25.32	25.36

[a] Elapsed Time (ET) to fit the model (in seconds)

the LSVC model, it provides the worst predictive capabilities compared with other tested models with a median AUC score of 73%. The DT model is the fastest model to train, requiring only 34.53 seconds and provides a median AUC score of 78%.

For better measuring the potential impact of the selected RF model, we consider the results of the last iteration of the RW (*U = 20*). Firstly, we selected the test set of the previous iteration (*U = 19*) to be used as the validation set for selecting the best threshold *D*, that represents an optimal interpretation of the predicted probabilities. This threshold is fixed and applied to the test set of the iteration *U = 20*, producing class labels. Figure 50 plots the ROC curve of RF model at iteration *U=19* and for *D* = 0.340, and Figure 51 shows the confusion matrix related with the RF model at iteration *U = 20* using the selected threshold.

Table 25: Comparison of the optimized ML models (best values in **bold**).

Models	**AUC**	**Training time (s)**	**Predict time (s)**
Random Forest (RF)	**0.90**[*]	943.99	0.079
Logistic Regression (LR)	0.76	57.47	0.072
Gradient-Boosted Tree (GBT)	0.82	13133.44	**0.045**
Linear Support Vector (LSVC)	0.73	1057.05	0.075
Multilayer Perceptron (MLP)	0.77	42.29	0.065
Decision Tree (DT)	0.78	**34.53**	0.070

[*] Statistically significant under paired comparison with LR, GBT, LSVC, MLP and DT.

Figure 52 shows the ranking (top 15) of feature importance for the RF model, as identified by the RF algorithm (Breiman, 2011). An interesting result is that the attribute *%Delays* was considered the most relevant. Moreover, another interesting result is that eight of the ten first relevant attributes were created

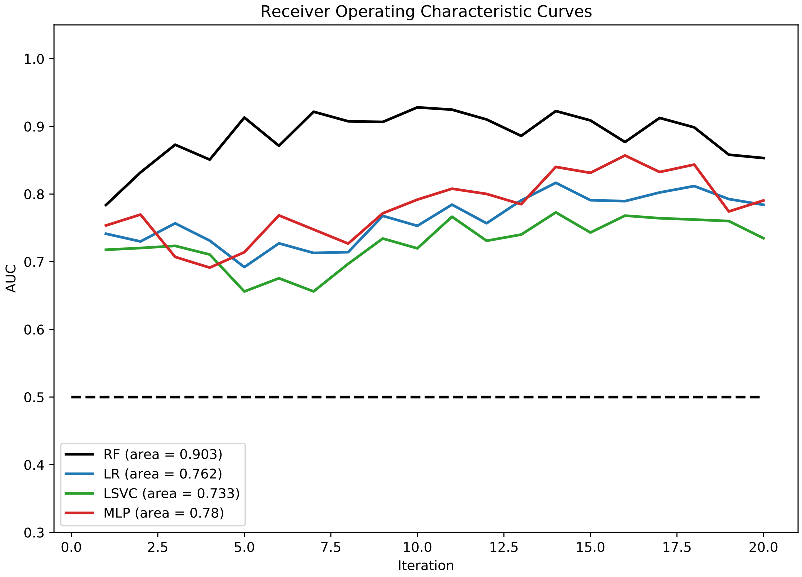
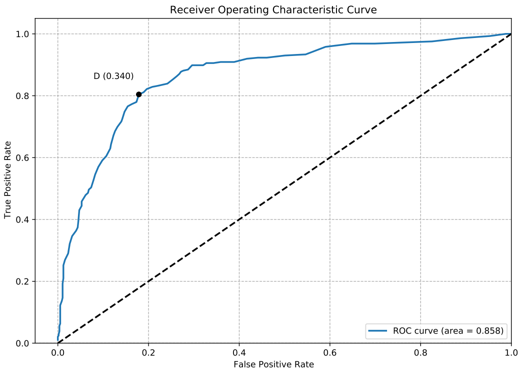
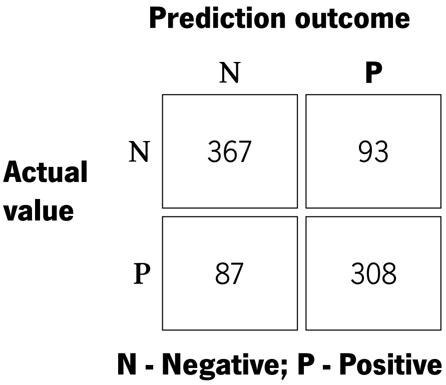


Figure 49: Evolution of AUC score over the RW iterations.





Prediction outcome

Actual value		N	P
	N	367	93
	P	87	308

N - Negative; P - Positive

Figure 51: Confusion matrix of the RF model at RW iteration *U = 20* when using *D* = 0.340.

on the feature engineering process using the domain expert knowledge (Section 4.2.5.3).

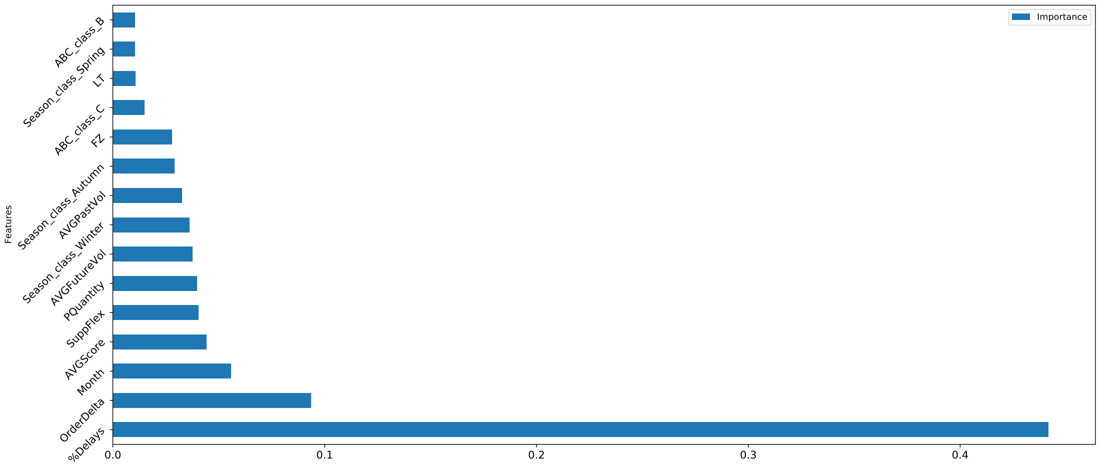


Figure 52: Feature importance for the RF model.

4.2.6.3 Misclassification impact on inventory management performance

A heuristic procedure was developed to calculate the costs associated with the misclassification of models.We used the knowledge of a business domain expert (as defined by the logistic planners) for calculating the special freight cost function. As described in Section 4.2.5.6, there are several factors that can induce variations in the special freight cost. Therefore, all costs considered were provided by domain experts and correspond to the average of costs charged in previous years by carriers for special freights.

Table 26 provides examples of special freight cost calculation for the case of misclassification of a positive instance as the negative one (FP) and also for misclassification of a negative instance as the positive (FN). These examples are related to six different orders of four raw materials in order to provide different cases of cost fluctuation.

The price of special freight for $C(1,0)$ tends to be higher than for $C(0,1)$ because in this case it is mandatory to react quickly to avoid stockouts, and normally the special freight service is contracted to be performed in one day (some exceptions can be extended for 2 days). In the case of misclassification of a negative instance as the positive (FN), special freight services are contracted for transit time in the range of 2-5 days.

We considered all the results of the RW iterations to measure the impact of misclassification on inventory management performance. Firstly, we selected the test set of each RW iteration as the validation set for selecting a particular threshold *D* (using the probability of positive class) that minimizes the costs. Then the selected threshold *D* in each RW iteration is fixed and applied to the subsequent iteration test results to determining the misclassification costs. For instance, in the first iteration is selected the threshold

Table 26: Examples of Misclassifying impact calculation.

Order nr.[a]	**RM nr.**[b]	**ABC**[c]	**Quantity**[d]	**Load unit**[e]	**Country**[f]	**C(1,0)**[g]	**C(0,1)**[h]
1	Raw material 1	A	8624	7	Malaysia	12.8k	12.3k
2			1568	1		8.0k	1.3k
3	Raw material 2	B	960	20	Czech Republic	11.3k	5.1k
4			48	1		568	416
5	Raw material 3	C	10000	1	National	41	417
6			5000	1		41	229
7	Raw material 4	C	3072	8	Germany	2.7k	472
8			384	1		781	148

[a] Order number;
[b] Raw material number;
[c] ABC classification of raw material;
[d] Quantity of raw material to be transported;
[e] Number of load unit for the raw material quantity;
[f] Country of dispatch;
[g] Cost of false negative (in monetary units);
[h] Cost of false positive (in monetary units).

D that provides minimum costs and then is used in the second iteration for determining the misclassification costs. In the second iteration the threshold *D* is also selected and applied to the third iteration, repeating this process until the last iteration (*U = 20*). Table 27 presents the total, average and median misclassification costs from all RW iterations of each tested models. These costs were calculated for a total of 4.580 orders, related with 109 raw materials from 19 suppliers. Furthermore, Table 28 describes the costs calculated for the selected threshold *D* in each RW iteration of the RF model.

Table 27: ML models cost (lowest costs in **bold**).

Models	**C(1,0)**[a]	**C(0,1)**[b]	**Costs**[c]		
			Total	**Average**	**Median**
Random Forest (RF)	17.6k	**9.1k**	**26.7k**	**1.3k**	**1.2k**
Logistic Regression (LR)	17.5k	15.8k	33.3k	1.6k	**1.2k**
Gradient-Boosted Tree (GBT)	**17.2k**	16.6k	33.9k	1.6k	1.6k
Multilayer Perceptron (MLP)	29.5k	11.6k	41.1k	2.0k	1.3k
Decision Tree (DT)	17.4k	14.7k	32.1k	1.6k	**1.2k**

[a] Cost of false negative (in monetary units);
[b] Cost of false positive (in monetary units);
[c] Misclassification Costs (in monetary units).

Regarding the results in Table 27, the RF model provides the lowest median cost (2.3k), followed by the GBT model (4.5k), LR, MLP and DT (5.5k). The RF model provides the best predictive capacities (see Table 25) and also provides the lowest total misclassification costs. On the other hand, the LR model provides the lowest costs related to the false-negative predictions for the selected thresholds, conversely to the GBT model that provides the higher costs. However, the GBT model provides lowest costs related to false-positive predictions, followed by the RF model. Overall, the RF model provides both better predictive capabilities and inventory performance.

Table 28: RF model cost for selected threshold *D* values.

RW Iteration ***U***	Threshold ***D***	C(1,0)[a]	C(0,1)[b]
1	0.34	0.0	201
2	0.34	884	1.2k
3	0.23	4.8k	0.0
4	0.7	310	1.9k
5	0.45	1.5k	557
6	0.47	600	616
7	0.52	150	833
8	0.17	1.3k	331
9	0.16	300	0.0
10	0.23	791	0.0
11	0.44	191	62
12	0.32	150	174
13	0.58	941	984
14	0.32	982	54
15	0.5	0.0	409
16	0.27	450	170
17	0.75	900	628
18	0.45	1k	300
19	0.47	900	601
20	0.21	1.2k	0.0

[a] Cost of false negative (in monetary units);
[b] Cost of false positive (in monetary units).

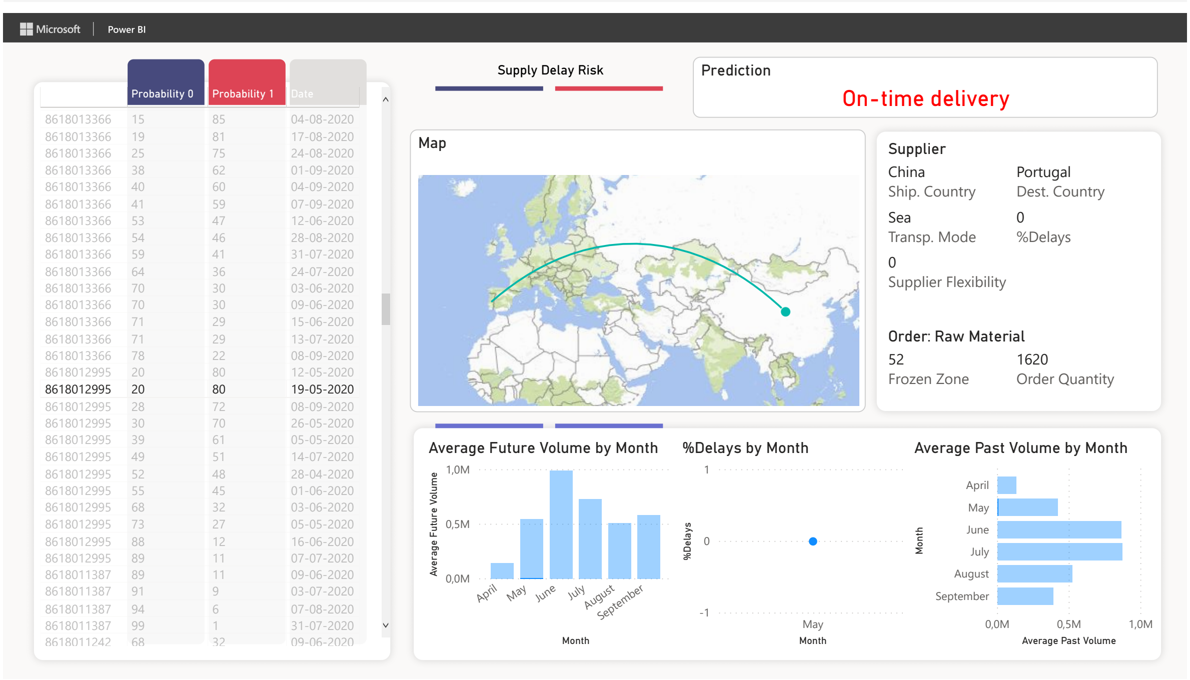
Table 29 shows that RF model predicts less FP and FN than other tested models. In our case, the $C(1,0)$ costs (related to FP) tend to be much higher due to the required quickly reaction to avoid stockouts. In particular, five of these FP predictions of the RF model consist of orders from outside Europe (Asia continent), where the price of one-day special freight is very high. Moreover, note that although the DT and LR models predict more FN and FP than the RF model, they provide equal median costs. This is due to the fact that the majority of the misclassification predictions of the DT and LR models are related to national orders where the special freight costs are relatively lower. The misclassification cost highly depends on the magnitude of the special freight cost established by the company for a given stock-out quantity, supplier and distance. As such, a reduced number of misclassifications could lead to significant costs for the organization if such misclassifications are made for highly-cost suppliers and raw materials. Following this reasoning, a classifier that performs poorly when predicting the delivery performance of low-cost orders might not produce high financial negative impacts to the organization.

Table 29: ML model predictions for selected threshold *D* values

Models	**TN**	**TP**	**FN**	**FP**	**TPR**	**FNR**
Random Forest (RF)	578	182	54	100	0.771	0.147
Logistic Regression (LR)	579	151	83	104	0.645	0.152
Gradient-Boosted Tree (GBT)	553	150	90	123	0.625	0.181
Multilayer Perceptron (MLP)	480	164	71	199	0.697	0.293
Decision Tree (DT)	575	168	72	104	0.700	0.153

In general, regarding logistics-related issues, it is relevant to consider a trade-off between inventory-related costs and Customer Service Level when selecting the best classifier. If the ultimate purpose of the company is to maintain high customer service level standards regardless of the inventory costs, then the RF is the best performing classifier. In sharp contrast, if the goal is to primarily minimize inventory costs regardless of the service level generated, then RF, LR and DT are the models that generate the lowest median inventory-related costs. In our particular case, we argue in favour of a compromise solution between inventory-related costs and prediction capability. In this sense, based on the results presented in Tables 25 and 27, the RF model can be considered the best classifier for our data in order to meet the aforementioned inventory management trade-off.

Figure 53 presents a dashboard developed using PowerBI software, with historical and predicted data related to the order deliveries. The predictions are the result of using the RF classifier. The displayed information can support the logistics planners' decision-making process regarding the procurement management.



	Probability 0	Probability 1	Date
8618013366	15	85	04-08-2020
8618013366	19	81	17-08-2020
8618013366	25	75	24-08-2020
8618013366	38	62	01-09-2020
8618013366	40	60	04-09-2020
8618013366	41	59	07-09-2020
8618013366	53	47	12-06-2020
8618013366	54	46	28-08-2020
8618013366	59	41	31-07-2020
8618013366	64	36	24-07-2020
8618013366	70	30	03-06-2020
8618013366	70	30	09-06-2020
8618013366	71	29	15-06-2020
8618013366	71	29	13-07-2020
8618013366	78	22	08-09-2020
8618012995	20	80	12-05-2020
8618012995	20	80	19-05-2020
8618012995	28	72	08-09-2020
8618012995	30	70	26-05-2020
8618012995	39	61	05-05-2020
8618012995	49	51	14-07-2020
8618012995	52	48	28-04-2020
8618012995	55	45	01-06-2020
8618012995	68	32	03-06-2020
8618012995	73	27	05-05-2020
8618012995	88	12	16-06-2020
8618012995	89	11	07-07-2020
8618011387	89	11	09-06-2020
8618011387	91	9	03-07-2020
8618011387	94	6	07-08-2020
8618011387	99	1	31-07-2020
8618011242	68	32	09-06-2020

Prediction
On-time delivery

Supplier

China Ship. Country	Portugal Dest. Country
Sea Transp. Mode	0 %Delays
0 Supplier Flexibility	

Order: Raw Material

52 Frozen Zone	1620 Order Quantity

Figure 53: PowerBI Dashboard: Supply delay risk prediction.

The left side of the dashboard shows information related to each order, such as the order number and the planned delivery date. Moreover, the probabilities of class 0 (delays) and 1 (on-time deliveries) from the classifier are also provided, giving the logistics planners information that can be useful to analyse the results of the predictions. It also shows a map illustrating the shipment country and the destination country of the selected order. Supplier information, such as the transportation mode and flexibility, and Raw material information, like the frozen zone, are also provided on the top right side of the dashboard. On the bottom of the dashboard, there is also information related to average future order volume, percentage of delays, and average past order volume by month. Note that such information agrees with the top 10 most important features for the RF classifier predictions, as illustrated in Figure 52.

4.2.7 Summary

Product quality, on-time delivery and manufacturing flexibility are the key factors for organizations to ensure a competitive advantage and improve their performance. On-time delivery is a crucial element for customer satisfaction and can have a direct impact on inventory levels, costs, and even in the company performance. There a new trend on using data analytics for predicting supplier disruptions in order to anticipating and managing future disruptions (Gonçalves et al., 2020).

In this context, we address the prediction of supply delay risk that aims to help the logistics planners of the organization in the decision-making process and consequently improving their efficiency, as well as avoid extra costs resulting from the occurrence of special freights. Therefore, we proposed a ML pipeline to build six different ML classification models supported by the proposed Big Data technological architecture. Firstly, we extracted logistics data from several sources and then stored in the Big data cluster using *Parquet* as storage file format. Afterwards, following the developed ML pipeline, we construct our dataset from the raw data, then we perform the data cleansing and create new information from the existing data (feature engineering).

We adopted a realistic and robust RW scheme to evaluate the classification models and the AUC as the evaluation metrics to compare the predictive performance of the following six learning classification models: Random Forest, Logistic Regression, Gradient-Boosted Tree, Linear Support Vector Machine, Multilayer Perceptron, and Decision Tree. Moreover, we also measure the impact of models misclassification on inventory performance. The results shows that the RF model provides the best prediction capabilities (median AUC of 90%), followed by the GBT, DT, MLP, LR and LSVC (73%). On the other hand, in terms of misclassification costs, the RF, LR and DT models provide lowest median cost (1.2k) followed by the MLP and GBT (1.6k) model. The RF model provides the best prediction capability and also lowest median costs in terms of misclassification costs. When considering the trade-off between inventory-related costs and service level, the RF model stands out as the best classifier. Therefore, we highlight the importance of measuring the cost of the model and its impact on the organization, as well as the importance of the business domain expert inputs to perform several tasks in a data mining project, mainly in the feature engineering tasks. We demonstrate that eight attributes of the top 10 relevant attributes were created on the feature engineering process using the domain expert knowledge.

The proposed ML framework provides several advantages, mainly in terms of operational and financial performance for the organization. We found that the operational performance, traduced by the improvement of the decision-making from the logistics planners, is the core advantage of our approach for the case-study company. The adoption of the proposed ML-based framework allowed the logistics planners to act proactively for possible delays and stabilize the inventory management process. In future work, we intend to consider a larger data set with more historical data collected by the organization. We also intend to extend this research study by creating new features that could improve the models predictions, create an automatic mechanism for the deployment and retraining of models, and assess the applicability of the proposed approach in other industries.

5 Supervised learning for estimating lead-time uncertainty

Summary: Uncertainty in supply lead time represents a core parameter that impacts both the supply chain performance and inventory parameters. Thus, improving supply lead time estimations can promote better estimations of safety stock. However, scarce attention has been given to supply lead time and in addiction, classical inventory models assume it to be deterministic. This assumptions is not realistic in real-world supply chain contexts due to unexpected events that can occur, which cause random delays. This chapter addresses the problem of estimating supply lead time for promoting not only better estimation of safety stocks, but also better logistics and transport management, production planning and management of capacities allocated to the production process at Bosch AE/P.

Chapter Table of Contents:

5.1 Introduction 136
5.1.1 Motivation 136
5.1.2 Research objectives and contributions 138
5.2 Related work 139
5.3 Problem formulation 142
5.4 Material and methods 143
5.4.1 Industrial case study 143
5.4.2 Big Data technological architecture 144
5.4.3 Machine Learning pipeline 145
5.4.4 Data Preparation 146
5.4.4.1 Logistics data 146
5.4.4.2 Feature engineering 148

5.4.5 Regression algorithms 150
5.4.5.1 Decision Tree (DT) regression 151
5.4.5.2 Random Forest (RF) regression 151
5.4.5.3 Linear Regression (LR) 152
5.4.5.4 Generalized Linear Model (GLM) 152
5.4.5.5 Gradient-Boosted Tree (GBT) regression 152
5.4.6 Feature preparation and hyperparameters tuning 153
5.4.6.1 Feature scaling 153
5.4.6.2 Feature selection 153
5.4.6.3 Hyperparameters tuning 153
5.4.7 Evaluation 154
5.4.7.1 Measuring model performance 155
5.4.7.2 Measuring model prediction bias 156
5.4.7.3 Safety stock estimation 156
5.5 Experiments and Results 157
5.5.1 Experimental and modeling setup 157
5.5.2 Preliminary experiments 158
5.5.3 Overall predictive performance 159
5.5.4 Safety stock performance 160
5.5.5 Practical & managerial implications 164
5.6 Summary 165

A machine learning strategy for estimating supply lead times towards improved safety stock dimensioning

Júlio Barros[1] · **João N.C. Gonçalves**[2] [3] · **Paulo Cortez**[4] · **M. Sameiro Carvalho**[2]

Submitted to a Journal

Abstract Supply lead time constitutes a core parameter in inventory control strategies and plays a critical role in supply chain performance. Although it is well-known that both uncertainty in demand and lead time can cause shortages or a surplus in inventory, scarce attention has been given to lead time variability by supply chain practitioners and researchers. In this paper, the main problem of interest is the estimation of supply lead time, which in the classical safety stock theory is often assumed as parametric, deterministic or constant. Motivated by the fact that such assumptions may not be realistic in many real-world supply chain contexts, we propose a novel IDSS that aims to predict supply lead time using a scalable technological Big Data architecture. Our main contribution relies on combining ML regression models and BD technologies for estimating lead time uncertainty. We focus on improving the estimation of lead time uncertainty to ultimately promote better safety stock estimations. We test our approach in real-world case study, using the estimated supply lead time as an input to dimension safety stocks. Under a realistic rolling window evaluation scheme, we compare the performance of five regression models, from which the RF model yielded competitive results with an average absolute error of 7.379. Numerical results show the benefits of the proposed IDSS in optimizing safety stock levels and inventory levels compared to those obtained by the case study company.

Keywords Supply chain risks, Lead time uncertainty, Safety stock, Machine Learning, Big Data.

5.1 Introduction

5.1.1 Motivation

SC is a dynamic, complex, and unique network of entities, processes, and resources (Chopra & Meindl, 2016; Simchi-Levi et al., 2000) involved in fulfilling the customer needs (Chopra & Meindl, 2016;

[1] ALGORITMI Research Centre/LASI, University of Minho, Guimarães 4800–058, Portugal.

[2] ALGORITMI Research Centre/LASI, Department of Production and Systems, University of Minho, Braga 4710–057, Portugal.

[3] Robert Bosch GmbH, Automotive Electronics Division, Logistics Section, Braga 4701–970, Portugal.

[4] ALGORITMI Research Centre/LASI, Department of Information Systems, University of Minho, Guimarães 4800–058, Portugal.

Council of Supply Chain Management Professional - CSCMP, 2013c). Within today's volatile, dynamic and competitive global market, supply chains are more exposed to several uncertainties and risks that cause disruptions and, ultimately, negative impacts on both customer satisfaction levels and supply chain-related costs (Er Kara et al., 2020; Hong, Lee, & Zhang, 2018; Z. Li et al., 2019). A recent example on this topic is the current pandemic of Coronavirus (SARS-CoV-2), which has been causing serious supply and demand issues in generalized supply chains. Indeed, this pandemic has made the supply chain even more exposed to these uncertainties and risks, making it more vulnerable to disruption events (Duong et al., 2022). However, it is well-known that managing such uncertainty factors and risks is a fundamental challenge for organizations, which has attracted the attention of many supply chain management researchers and practitioners over time. Despite the fact that there are various uncertainties and risks in the field of SC (e.g., supply Lead-time (LT) uncertainty, customer demand uncertainty, yield uncertainty and order crossover), Heydari et al. (2009) argued that the most important ones relate to customer demand and supply LT uncertainty. Supply LT is broadly defined as the time between the placement of an order and its receipt (Chaharsooghi & Heydari, 2010; Heydari et al., 2009; Juez et al., 2010; Singh & Soni, 2019), and according to Chaharsooghi and Heydari (2010), it comprises several components such as information delays in receiving orders by the downstream players, time of order processing and transportation time.

Conceptually, uncertainty in supply LT constitutes a core parameter that varies and affects the SC performance and inventory parameters (Chaharsooghi & Heydari, 2010; B. Dey et al., 2021; Heydari et al., 2009; Z. Li et al., 2019). However, scare attention has been given to supply LT management and researchers have paid more attention to demand uncertainty (Dolgui & Prodhon, 2007; Heydari et al., 2009). A possible explanation for this may be the fact that supply LT can be influenced by uncontrollable factors (Heydari et al., 2009), such as strikes, weather conditions, customs issues, as well by the bullwhip effect from downstream SC players. In fact, although research on stochastic lead times has been ongoing for decades (see, e.g., Scarf, 1958; Silver et al., 2016; Zipkin, 2000), M. Louly and Dolgui (2009) argued that *"stochastic lead time models are less developed and more complex"*. We note that previous research (Bandaly et al., 2016; W. Chang & Lin, 2019; Chatfield et al., 2004; Duc et al., 2008; Hayya et al., 2011; X. He et al., 2011; Heydari et al., 2009; Michna et al., 2018) have assessed the impact of supply LT on supply chain performance. However, most of research works have been assuming supply LT as deterministic and do not exploit non-parametric approaches that leverage the interaction between different variables with potential impact on a suppliers' LT.

Buffering techniques such as Safety Stock (SS) are adopted by organizations to cover both demand and LT uncertainties in order to achieve the promised service level to the customers and prevent stock-outs (C. A. Chang, 1985; Jung et al., 2008). Safety stock plays a crucial role in maintaining the balance between excess inventory and lost sales, which leads to better SC performance. There are a significant literature contributions pertaining the SS dimensioning and typically under the assumptions of stationary demand (following a normal/Gaussian distribution) and stationary LT (Kanet et al., 2010). Indeed, LT is treated as deterministic or constant for the most of inventory models, which is not realistic for major supply chain environments due to the unexpected events that can occur causing random delays (B. Dey et al., 2021). Research study of Chopra et al. (2004) assessed the effects of LT on SS, and recently W. Chung et al. (2018) investigated the effects of LT uncertainties (replenishment lead-time and call-off lead-time uncertainties) and SS on logistics performance in a JIT supply chain. Previous work of Barros et al. (2021) stated that the safety stock research problem has been gained increased attention from researchers and practitioners since 2007 until now. Moreover, they argued that optimization and simulation-based optimization are the two main approaches adopted in the literature to tackle safety stock problems. On the other hand, Gonçalves et al. (2020) highlighted the importance of exploring data-driven approaches such as, BDA for the enhancement of logistics decision-making processes, including the management of safety stocks. Indeed, unlike conventional statistical approaches, the use of data-driven approaches allows, for instance, the inclusion of other variables that impact on the dynamics of a supplier LT. On the other hand, if properly selected, such variables also allow capturing recent changes in supplier response patterns with greater flexibility. Despite the advantages of using BDA to model stochastic dynamics, only a few works have considered the existence of LT variability issues (see, e.g., Abdel-Malek et al., 2005; Chopra et al., 2004; Digiesi et al., 2013; Disney et al., 2016; Kanet et al., 2010; M. Louly & Dolgui, 2009; Ruiz-Torres & Mahmoodi, 2010; Saad, Perez, & Alvarado, 2017; Talluri et al., 2004).

Our research is essentially motivated by: (i) the lack of data-driven approaches for modeling supply LT; (ii) the lack of research works focusing on providing dynamic models that consider the knowledge of future volatility of parameters for safety stock estimation (Barros et al., 2021); (iii) the lack of studies considering stochastic LT in general MRP inventory systems (Barros et al., 2021); (iv) the lack of reported real-world case studies applying safety stock dimensioning methods to supply chain with multiple products and with a large amount of data (Gonçalves et al., 2020); and (v) the importance of improving safety stock estimation at upstream stages of SC using BDA approaches (Gonçalves et al., 2020).

5.1.2 Research objectives and contributions

This paper aims to propose an IDSS that combines ML and BD techniques to support both supply risk and inventory management. This IDSS includes a ML approach for predicting LT uncertainty supported by a scalable technological Big Data architecture proposed in N. Silva et al. (2021). Moreover, the proposed approach allows the enhancement of SS levels computation (through a traditional and well-known method) considering the outcomes from the proposed ML approach. We explore five different ML regressions models (Random Forest, Linear Regression, Generalized Linear regression Model, Gradient-Boosted regression and Decision Tree regression). Indeed, we applied and evaluated this proposed approach in a real-world multinational automotive electronics organization - Bosch AE/P, Portugal. We evaluate the ML regression models not only in terms of predictive power and bias but also in terms of inventory-related costs.

Currently, there is a static approach in the case study company to estimate safety stocks. This approach has been refined and improved by logistics planners over the years based on their experiences and observations rather than in technically-based approaches. The proposed DSS, in addition to enabling the prediction of LT uncertainty, also provides a systematic approach to determine safety stock minimizing the holding inventory costs while attending to a certain service level. Overall, the proposed approach proves to be valuable to the organization in terms of supporting the decision-making process of the logistics planners.

The main contributions of this work can be summarized as follows:

- we propose an ML pipeline for predicting stochastic LT supported by a scalable technological Big Data architecture proposed in N. Silva et al. (2021);
- we compare the predictive power of five ML regression models (Random Forest, Linear Regression, Generalized Linear regression Model, Gradient-Boosted Tree regression and Decision Tree regression) using real-world supply chain data;
- we adopt a robust RW models evaluation procedure for producing a set of training and test modeling iterations in order to simulate a realistic supply chain planning environment;
- we evaluate the performance of ML models using statistical and supply chain inventory metrics;
- we propose a dynamic approach for estimating safety stocks under dynamic manufacturer's demand and stochastic LT, in order to promote the minimization of inventory-related costs while maintaining a proper customer service level;
- we propose a flexible approach that allows the company to evaluate the impact of a given safety stock level in terms of inventory-related costs.

The rest of the paper is organized as follows. Section 5.2 provides an overview of literature contributions related to LT modeling using BDA techniques. Section 5.3 describes the problem under consideration. In Section 5.4, we provide details on the proposed ML framework. Section 5.5 outlines the experimental and modeling setup, and obtained results from the proposed IDSS. Finally, Section 5.6 highlights the main results and conclusions.

5.2 Related work

In the past, scarce attention was given to LT management by researchers and practitioners. A large majority of traditional inventory models have been assumed LT to be deterministic or Gaussian distributed stationary. While providing greater simplicity to the modeling processes, the first assumption implies that the safety stock derived from it may be unable to handle the irregular and ever-changing behavior naturally inherent to the operating processes of most suppliers. Consequently, this may lead to potential inventory shortages and damage to the customer service level. On the other hand, the assumption of Gaussian distributed stationary LT is not realistic in practice, mainly due to the complexity of real-world SCs that makes them more susceptible to unexpected events that cause supply delays.

Lead-time uncertainty has increasingly attracted the attention of researchers and practitioners due to its effects on supply chain performance measures, such as inventory costs, bullwhip effect, and product availability (Chaharsooghi & Heydari, 2010; Heydari et al., 2009). Mitigating these effects leads to a reduction in supply chain response time, creating competitive advantages for companies (Gunasekaran et al., 2001). Indeed, several techniques have been employed for modeling LT, including Operation Research (OR) and learning-based techniques. OR techniques such as optimization and simulation are the most used, whereas LR, Artificial Neural Network (ANN) and DT are the typical (statistical/machine) learning strategies adopted (Burggräf et al., 2020; Öztürk et al., 2006).

Lead-time related problems can be classified into: order/supply LT and operation LT. The order/supply LT represents the time between placing an order and its reception. On the other hand, operation LT is divided into interoperation (includes the wait time and transportation time) and operation time (includes setup time and actual processing time) (Burggräf et al., 2020). By searching upon the literature on LT estimation (see Table 30), we observed that there is a strong focus on modeling the operation LT. For instance, Öztürk et al. (2006) explored two data mining algorithms (DT and LR) in order to estimate the manufacturing LT in MTO manufacturing environment composed of three different job shops (SHOP-V, SHOP-A and SHOP-I). All data used was generated through computer simulation. Juez et al. (2010) proposed an ML model to predict the manufacturing times (manufacturing LT) of different metallic components of aerospace engines and uses the Root Mean Squared Error (RMSE) metrics to measure predictive power. Gyulai, Pfeiffer, Nick, et al. (2018) proposed also a ML approach to determine manufacturing LT of a flow-show manufacturing environment. They explored the LR, RF and SVM algorithms. Then, the proposed ML-based approach is compared with the Little's law analytical LT prediction method. The authors concluded that the proposed approach can outperform the analytical method regarding the LT prediction due to the dynamics of the system and the efficient consideration of the job features.

To the best of our knowledge, scarce attention has been given to the development of learning-based techniques to estimate order/supply LT. A recent exception is the work of Singh and Soni, 2019, exploring several ML algorithms to predict the LT for a JIT manufacturing environment from a restaurant, using statistical-based error metrics to evaluate the model's predictive power. In contrast with the work of

Table 30: Chronological scientific works for modeling LT using BDA techniques

Study	**T**[a]	A[b]	**M**[c]	**D**[d]	**E**[e]	**P**[f]	**R**[g]	**L**[h]	**S**[i]
Öztürk et al. (2006)	ML	DT, LR	MSE, R^2, CV, AAE, RE	No	HO	-	38k	OP	MTO
Juez et al. (2010)	ML	SVM	RMSE	Yes	HO	-	634	OP	-
Gyulai, Pfeiffer, Nick, et al. (2018)	ML	LR, SVR, RF	NRMSE	Yes	-	-	5k	OP	FS
Gyulai, Pfeiffer, Bergmann, and Gallina (2018)	ML, S	LR, RF	NRMSE	No	-	-	-	OP	PPC
Lingitz et al. (2018)	ML	LR, RR, LaR, ANN, SVM, MARS, KNN, ANN	MAE, MAPE, MSE, RMSE, NRMSE	Yes	HO, CV	2y	18.5k	OP	PPS
Singh and Soni (2019)	ML	LR, RR, LaR, DT, RF, KNN	MAE, MSE, RMSE	Yes	-	-	-	O	JIT
Lim et al. (2019)	ML	SVM, ANN, RF	ACC, TPR, FM	Yes	CV	1y	-	OP	ATO
Hathikal et al. (2020)	ML	LoR, SVM, DT, KNN	ACC, TPR, PPV	Yes	CV	13m	-	OP	-
Jeong et al. (2020)	ML	DT, ANN, MLP	MAPE, MAE, RMSE, RMSLE	Yes	HO	6y	118.7k	OP	MTO
Bender and Ovtcharova (2021)	ML	GBM	RMSE	No	-	-	250k	OP	MTO
This work	ML, BD	LR, DT, RF, GLM, GBT	MAE, MSE, RMSE, R^2, AUREC	Yes	RW	2.6y	32k	O	MTO

[a] Modeling **T**echnique: ML - Machine Learning, S - Simulation, BD - Big Data.

[b] **A**lgorithm: SVM - Support Vector Machine, RF - Random Forest, LR - Linear Regression, LaR - Lasso Regression, LoR - Logistic Regression, RR - Ridge Regression, KNN - K-Nearest Neighbors, ANN - Artificial Neural Network, MLP - Multi-layer Perceptron MARS - Multivariate Adaptive Regression, GBT - Gradient-Boosted Tree, GLM - Generalized Linear Models.

[c] **M**etrics: RMSE - Root Mean Square Error, NRMSE - Normalized Root Mean Square Error, MAE - Mean Absolute Error, MSE - Mean Squared Error, MAPE - Mean Absolute Percentage Error, RMSLE - Root Mean Squared Logarithmic Error, AUREC - Area under the Regression Error Characteristic curve, R^2 - Coefficient of determination, TPR - True Positive Rate, PPV - Positive Predictive Value, ACC - Accuracy, F1 - F1 score, AAE - Average Absolute Error, CV - Coefficient of Variation, RE - Relative Error.

[d] Empirical **D**ata.

[e] Model **E**valuation: HO - Hold-out, CV - Cross-validation, RW - Rolling Window.

[f] Data **P**eriod: m - month, y - years.

[g] Data **R**ecords: k - thousands of records.

[h] Type of **L**ead-time: O - Order lead-time, OP - Operation lead-time.

[i] **S**upply chain environment: FS - Flow-shop, JIT - Just-in-time, ATO - Assembly-to-order, PPC - Production Planning Control, PPS - Production Planning and Scheduling.

Singh and Soni, 2019, we investigate the performance of ML-based models not only in terms of predictive power but also in terms of prediction bias and inventory-related costs. Moreover, to the best of our knowledge, our work is the first in the supply chain management literature combining ML and BD to estimate supply LT in a real-world multinational automotive electronics supply chain (Bosch AE/P) operating with an MRP inventory control system. We build on the scalable technological BD architecture proposed in N. Silva et al., 2021. Our IDSS builds on a BDW that stores, integrates, and provides real data over time. We consider the LT estimations provided by our ML-based models for dimensioning safety stocks purposes. In such a setting, safety stocks assume that manufacturer's demand, resulting from the dependent demand of end customers, is deterministic and known in advance but dynamic and time-varying across the entire planning horizon. In other words, the manufacturer's demand is obtained through the case-study company's ERP-MRP, which receives (deterministically) the time-varying dependent demand for finished products to further run the MRP for components. Importantly, we evaluate the impacts of our approach, providing a comparison with the current strategy used by Bosch AE/P, in terms of SS estimations and inventory holding costs. Our IDSS derives safety stock estimations adapted to the recent supply dynamics, as opposed to the static and experience-based approach currently used by the company.

5.3 Problem formulation

We consider a standard supply chain composed by three players (supplier, manufacturer and customer) establishing exchanges of information and materials over time, as shown in Fig. 54. In order to cope with upstream and downstream variations in the supply chain, the manufacturer dimensions safety stocks for multiple components. For this purpose, classical inventory management theory states that one must correctly characterize and further model the demand and supply processes in order to promote buffer stocks that minimize holding costs while maintaining the desired service levels (Silver et al., 2016). In this paper, we focus our attention on the problem of estimating supply LT, defined here as the length of the interval between the placement of the supplier order and the time it reaches to the manufacturer. In this way, we consider not only the supply transit time but also the time required to produce the quantity requested by the manufacturer from the supplier. Note that the importance of promoting better supply LT estimates goes far beyond the problem of dimensioning safety stocks, to the extent that it also has significant impacts on production planning and leveling, in the management of capacities allocated to the production process, and in the logistics and transport management.

Let $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^{n}$ be a training set comprising n observations where $\mathbf{x} \subseteq \mathbb{R}^p$ denotes a p-dimensional vector of input variables. For each training observation $\mathbf{x} \in \mathcal{T}$, there exists a unique corresponding outcome $y \in Y \subseteq \mathbb{R}$. Hence, a supervised learning training set $\mathcal{D}$ can be formulated in the form of multiple input-output relationships as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \subseteq \mathbb{R}^p \times \mathbb{R}$. The main purpose of this work is to construct a regression framework that takes advantage of a subset input variables (embedded in $\mathbf{x}$) in order to promote better estimates of the supply LT (y). By means of a regression model R, characterized

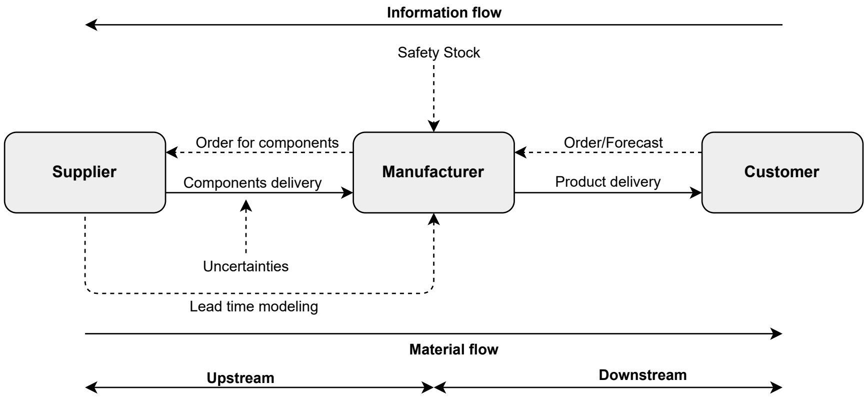


Figure 54: Illustration of a classical three-stage supply chain structure.

by an optimal hypothesis, we intend to map linear/nonlinear relationships that coexist between the p-dimensional input vector $\mathbf{x}$ and the corresponding output y.

5.4 Material and methods

5.4.1 Industrial case study

This work was developed in the Logistics Department of Bosch AE/P. This multinational automotive electronics industry is characterized by a complex supply chain topology, with suppliers and customers spread all around the world, as depicted in Figure 55. The Logistic Department deals with a total of 7781 raw materials from 444 suppliers located in more than 30 countries, which impacts the production of about 1100 finished products. The majority of suppliers (48%) are located in the Asian continent, followed by the European continent (43%, in which 8% are Portuguese suppliers) and the American continent (only 1%).

Progressively, the automotive industry has incorporated the JIT concept, which is concerned with demand-driven production in order to reduce overall waste, particularly inventory levels. While it is true that the adoption of this philosophy allows the reduction of inventory on-hand, it may also result in higher vulnerabilities to uncertainty phenomena. Managing these uncertainties and risks becomes a fundamental challenge for the organization. In this way, the company adopts MRP buffering techniques such as safety stocks in order to hedge against inventory shortages induced by supply and demand uncertainties. Establishing an appropriate level of safety stocks is a complex task since a low safety stock can lead to shortages and a high safety stock can lead to overstocking, and thereby higher and unnecessary holding and warehousing costs.

The current strategy used at the case study company to estimate the safety stock is based on the logistics planner's experience. In this process, several important logistics criteria are considered, including: ABC classification by ordering volume and unit cost, supplier origin and transportation mode, special freight lead time, and effects of planning time fence. However, although these criteria are proven to have

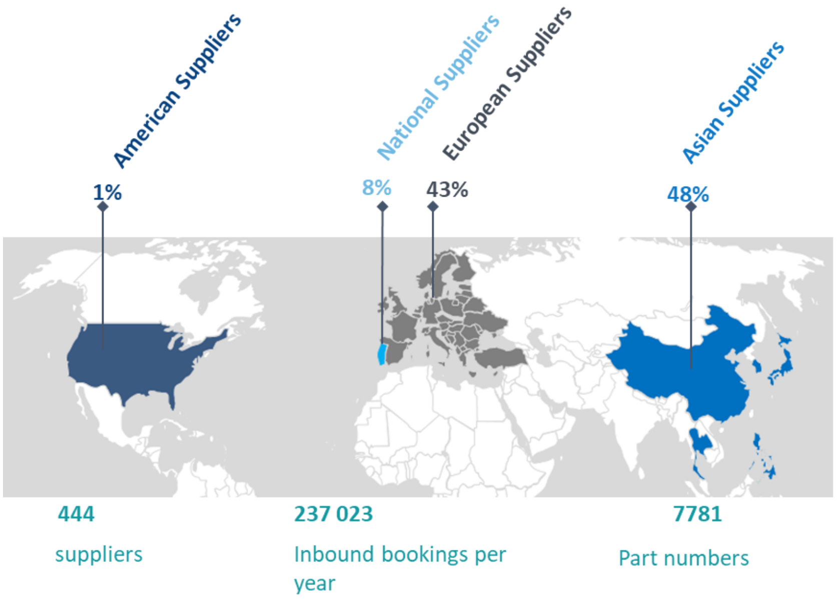


Figure 55: Geographic distribution of the Bosch AE/P suppliers.

a direct impact on inventory management, the company lacks a dynamic and systematic strategy that allows it to aggregate all these factors in an efficient way. On the other hand, it is noteworthy that this gut-feeling process based on past experience is quite error-prone and requires a lot of knowledge on the business process. Hence, considering the case study company as practical motivation, we intend to promote better estimations of safety stocks by improving supply LT estimations. For that we propose a IDSS that combines ML and BDA technologies. The details of our approach are detailed in the subsequent sections.

5.4.2 Big Data technological architecture

This section describes the Big Data technological architecture that supports our modeling approach to estimate supply LT. Such architecture was previously introduced in N. Silva et al. (2021) and includes three main layers described as follows:

A. **Sources**: consists of SQL databases or other data repositories from Bosch AE/P.

B. **Big Data Infrastructure**: an Hadoop-based ecosystem comprising the following tools:

   1. **Sqoop** - used for data ingestion from the selected sources. Collected data from data sources are stored in an HDFS in *Parquet*[1] format, enabling compatibility with Apache Spark and Impala, as well as a high degree of efficiency in large-scale queries (Baranowski et al., 2015).

[1] https://parquet.apache.org/

2. **Apache Spark** - a multi-language engine used for data cleansing, transformation and development of ML models. This engine was adopted primarily due to its scalability, simplicity and easy integration with other BD tools.
3. **Hive**[2] - a distributed data warehouse system used to store and manage large-scale datasets. It includes the Hive Metastore where the metadata, namely schema and other model statistics, are stored (Thusoo et al., 2009).
4. **Impala** - a native analytic database for Apache Hadoop that we have used to query data from Hive and provide a connection to the Power BI tool. The choice of impala was motivated by its query execution performance when compared with other SQL-on-Hadoop systems, namely Spark, Drill, Presto or Hive (M. Santos et al., 2017; N. Silva et al., 2021).

C. **Data visualization**: module that enables to create data visualizations through Power BI dashboards. Note that Power BI is a powerful, popular, easy to use tool, considered to be a top business intelligence platform[3].

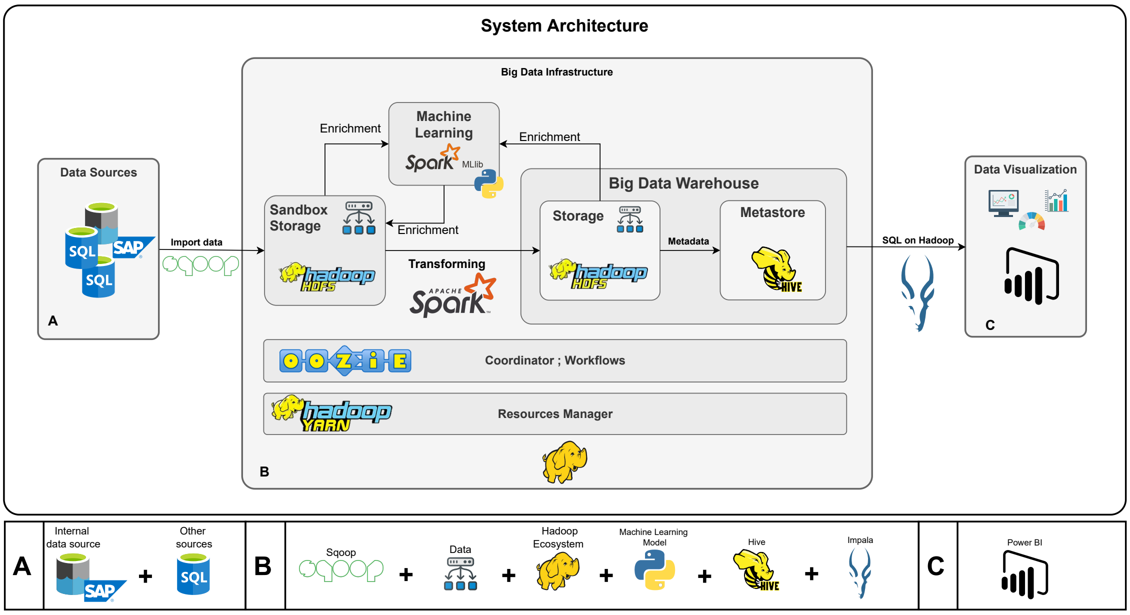


Figure 56: Big Data technological architecture proposed in N. Silva et al. (2021)

5.4.3 Machine Learning pipeline

The proposed approach appears summarized in Figure 57 and relies on three core data mining phases. The first concerns the data preparation, where we perform several activities related to data pre-processing, including data cleaning, feature selection and feature standardization (Section 4.2.5.3).

[2] https://hive.apache.org/

[3] https://info.microsoft.com/ww-landing-2022-gartner-mq-report-on-bi-and-analytics-platforms.html

The second phase relates to the modeling and evaluation, where we build the predictive models and evaluate their performance (Sections 4.2.5.4, 4.2.5.5 and 4.2.5.6). The third phase is the deployment, where we run the predictive model in a Big Data Hadoop Cluster. In the whole process, we create a dynamic workflow that, firstly, trains the selected ML model, then creates a pickle of this model, and lastly, saves it into a selected HDFS folder. A pickle is a Python module used for serializing and de-serializing object structures. Oftentimes, this module is used to save ML models and, in a later stage, generate new predictions without re-training them. All the derived predictions are then stored in Hive external tables of the Big Data Warehouse (see Fig. 56).

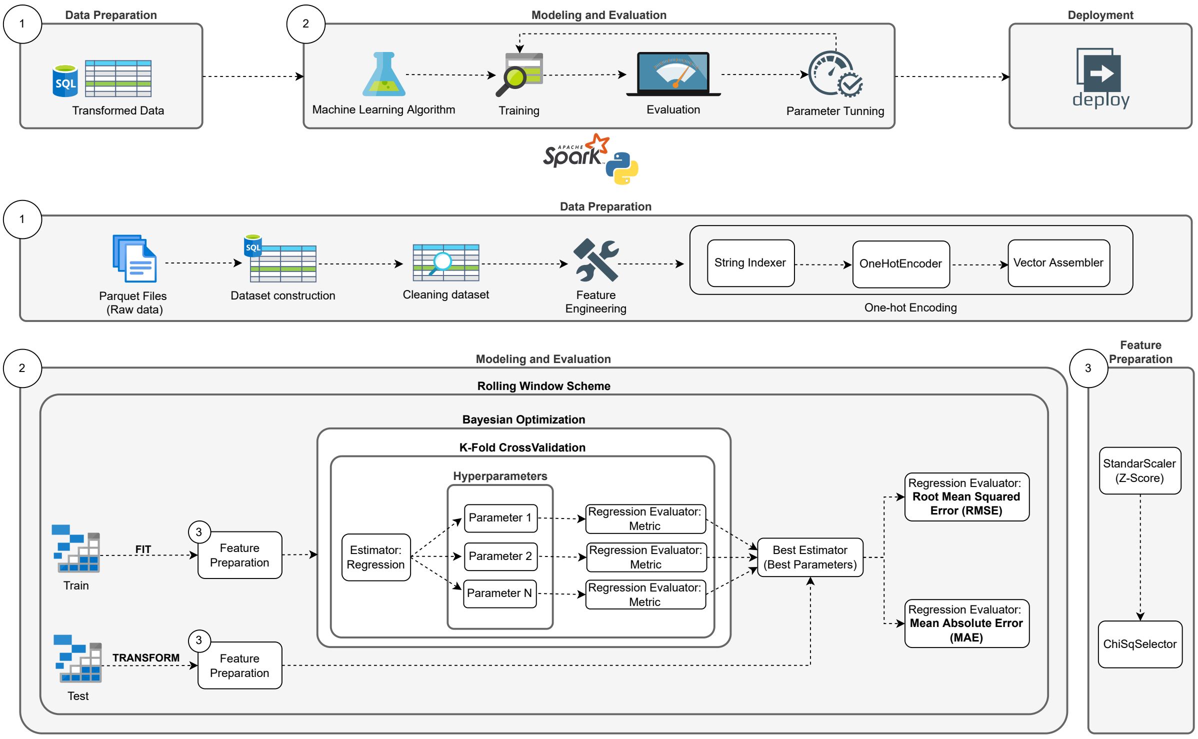


Figure 57: Machine learning pipeline.

5.4.4 Data Preparation

5.4.4.1 Logistics data

This research uses proprietary data from the Logistics Department of the multinational automotive electronics industry (Bosch AE/P), from February 2019 to August 2021. The collected data were retrieved from the BDW proposed in N. Silva et al. (2021) in agreement with the business domain expert knowledge (logistic planners). Table 19 presents a description of the raw logistics information that served as basis for the feature engineering process, from which the input features described in Section 5.4.4.2 were created.

Table 31: Summary of the collected logistics data attributes.

Context	**Attribute**	**Description (format, {examples})**
Orders	Order date	Order placed date (date, {'20190320'})
	Order quantity	Quantity of ordered item (number, {1800, 10000})
	Quantity delivered	Quantity delivered (number, {1500, 8000})
	Delivery date (Planned)	Planned date to deliver the order (date, {'20190802'})
	Delivery date (Real)	Date of order delivery (date, {'20200302'})
	Transport mode	Type of transportation (text, {'Sea', 'Air', 'Road'})
	Forwarder	Name of delivering carrier (text, {'Company X'})
Materials	Material ID	Raw material code (text, {'00126523', 'FH123201'})
	Planning time fence	Frozen zone in days at supplier (number, {2, 8})
	ABC classification	Classification of raw material determined by volume of sales and price (text, {'A', 'B', 'C'})
	Planning calendar	Delivery cycle code of supplier (text, {LSP, ERS})
Suppliers	Number	Code of supplier (text, {'0685RF5T'})
	Name	Legal name of the supplier (text, {'Company Y'})
	Location (country)	Location country of supplier (text, {'Portugal'})
	Location (region)	Region of the supplier location (text, {'Europe'})
Plant	Plant number	Plant code (text, {'8051', '9245'})
	Name	Plant name (text, {'Plant 1'})
	Plant location	Country where plant is located (text, {'Portugal'})
Calendar	Holidays	Number of holidays during the planned LT from the country of dispatch (number, {1, 4})
	Planning delivery calendar	Planned delivery calendar code (text, '1A02', 'RS42')
Target	Lead time	Supply lead time (number, {2, 19})

We have conducted data cleansing activities to remove all missing and duplicated data. Afterwards, we obtained a data set of 32k orders related to 2.6k different raw materials, from a total of 174 suppliers. While performing the EDA, we found that the distribution of our target output (LT) is skewed (right-skewed), as shown in Fig. 58. It is well-known that the tail region in skewed data can act as an outlier in statistical models. Therefore, we applied the logarithmic-based transformation $log(y+1)$ to the target output in order to alleviate the skewness of the distribution (Hastie et al., 2009). The target in the regression thus becomes $log(y+1)$, and thereby all model predictions should be post-processed by using the respective inverse function. Note that we have chosen the $log(y+1)$ transformation rather than $log(y)$ due to the existence of many zeros-valued observations in our target output (y), which makes the latter transformation unfeasible.

Lastly, we have also applied the well-known one-hot encoding technique to convert the categorical features into numerical ones, making it possible to use models that require numerical inputs as a precondition. For each categorical input in the dataset, this technique creates a new binary dummy variable for each unique category value therein represented (James et al., 2021).

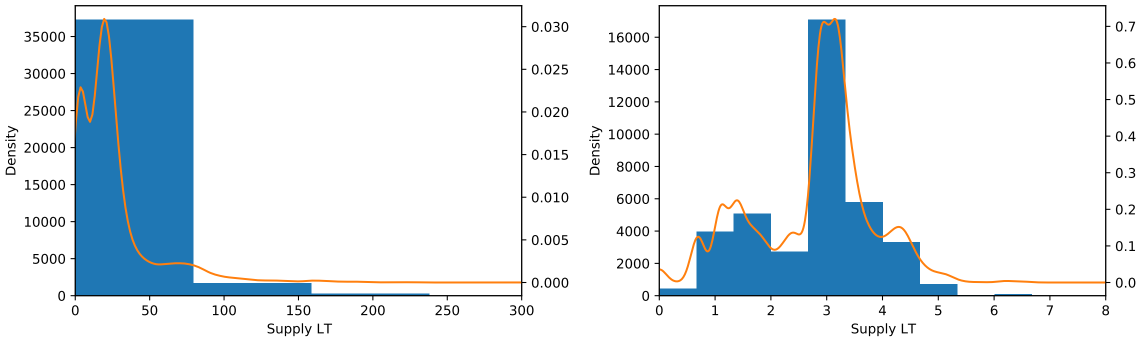


Figure 58: (left) LT and (right) $log(LT+1)$ distributions with Kernel Density Estimates (KDE)

5.4.4.2 Feature engineering

Domingos (2012) states that feature engineering represents a crucial process regarding any ML task, which aims to create a set of new features from raw data in order to improve the predictive power of ML models. Following previous research (Brintrup et al., 2020), this section introduces all the features considered during the learning processes of the suppliers' delivery time dynamics. Such dynamics, embedded in the multidimensional vector **x**, may one hand positively/negatively affect the ability of a supplier to deliver a given order on time and, on the other hand, explain increases/decreases in the supply lead time throughout the product life-cycle. The constructed features are described as follows:

- **Historical percentage of supplier delays**: The past recent information related to delays caused by the supplier may reveal some trends regarding the increase/decrease of supply lead time. In fact, suppliers associated with frequent late deliveries are more likely to delay future deliveries. Hence, at the time of placing a given order t, we compute the average percentage of delays, induced by the supplier, over the past n orders:

$$SupplierDelays(t)^{(n)} = \frac{100}{n}\sum_{i=1}^{n} 1_{\{A_{t-i}>R_{t-i}\}} \tag{5.1}$$

  where A_{t-i} and R_{t-i} are the actual and required receipt dates for the past supplier order $t-i$, and 1_A denotes a logical function (it equals 1 if A is true and 0 otherwise). In this formulation, t represents the index of the supplier order under prediction. A natural consequence of applying the expression (5.1) is that the first n instances of the data are dropped due to the use of time lags of order n.

- **Historical percentage of forwarder delays**: Variations in the lead time of a certain order may not necessarily be related to factors related to the supplier level but, for example, related to the forwarder. By way of example, issues like accidents and customs clearance problems impact strongly in the ability to promote just-in-time deliveries. Thus, similarly to the preceding variable, we also opted to include the historical percentage of delays motivated by logistics issues at the forwarder and not the supplier.

- **Average past/future order volume**: We consider that the amount of material requested to the supplier can influence its production capacity to an extent that could induce variations in the delivery time of a particular order quantity. In order words, the volatility, in terms of quantity, of manufacturer's orders to the suppliers may influence their ability to cope with the manufacturer's requests. Following this reasoning, we select the level of manufacturer's demand volume over the last n orders as a way to reflect the past pressure exerted on the supplier:

$$AVGPastVol(t)^{(n)} = \frac{1}{n}\sum_{i=1}^{n} O_{t-i} \tag{5.2}$$

  Likewise, we have also extended the previous formulation to consider the average future order volume, planned today for the future n orders (Eq. 5.3), in the sense that large future order quantities can force the supplier to reschedule current production and capacity which, in turn, can compromise the supply lead times of short-term orders:

$$AVGFutureVol(t)^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{O}_{t+i}\ , \tag{5.3}$$

  where $\widehat{O}_{t+i}$ is the ith forecasted order.

- **Short-term supplier flexibility**: Following the work of Brintrup et al. (2020), we have considered the supplier flexibility as one of our model features. In our work, we define the supplier flexibility as the first-order difference between the quantity to be ordered and the last ordered quantity, using the following expression:

$$ShortSuppFlex(t) = \nabla O_t = O_t - O_{t-1}\ . \tag{5.4}$$

  With this formulation, we expect that by applying learning algorithms it should be possible to correlate short-term variations in the quantity to be ordered with increases/decreases in supply lead times.

- **Mid-term supplier flexibility**: We extend the previous formulation over the past n orders in order to account for mid-term variations, rather than just short-term dynamics:

$$MidSuppFlex(t)^{(n)} = O_t - \frac{1}{n}\sum_{i=1}^{n} O_{t-i} \tag{5.5}$$

  In this sense, we can explicitly account for events where order variations were high in the short-term but smoothed in the mid-term, and vice-versa, and correlate them with past supply lead time increases/decreases.

- **On-time delivery moving average**: For each record of a placed order, it is possible to determine whether the order arrived on the scheduled date set by the producer. Let X be a binary variable that for a given placed order t takes the value 0 if the order did not arrive on the scheduled date and 1 otherwise. For each new order entry in the future, we calculate the nth order moving average (MA) of X aiming to try to smooth its behavior over time, and to ultimately measure the trend of a given supplier to incur in delays:

$$MA_t(n) = \frac{1}{n} \sum_{i=t-n}^{t-1} X_i \tag{5.6}$$

- **Number of non-working days**: Another variable that can influence the length of the lead time is the number of non-working days (prevailing in the supplier's country of origin) from the moment that the order is placed at the supplier by the producer until the planned date of its receipt.

- **Order Frequency**: Consists of determining the frequency of orders to a specific supplier. Note that when compared to suppliers with shorter replenishment lead times, a supplier with more time-spaced delivery frequencies can be expected to be more prone to delays in order deliveries, due to the greater number of uncertainty factors that exist between itself and the producer.

Throughout the empirical evaluations considered in this paper, we have considered $n = 10$ orders. This assumption was derived from a brainstorming meeting with business expert domains, where it was considered a reasonable time frame to capture relevant supplier dynamics. In any case, it should be noted that this value can, of course, vary depending on the application context at hand.

In addition to the abovementioned variables, we also consider some fundamental logistics information related to the component to be ordered, including its: (i) order quantity, (ii) transport mode, (iii) planning time fence, (iv) ABC Classication, (v) planning calendar, (vi) supplier location, (vii) supplier region, (viii) plant location, (ix) number of holidays, (x) planned delivery calendar, (xi) season of the year, (xii) month of the supply order. This results in a set of 21 features for modeling purposes plus the target variable (LT).

5.4.5 Regression algorithms

During several meetings with the company business experts, we have conducted a controlled set of exploratory visual inspection analyses that confirmed the existence of several non-linear relationships between the dependent variable the independent variables. These findings motivated the use of three nonlinear regression algorithms (DT, RF and GBT). Yet, we have also chosen to include two linear models (Linear Regression (LR) and Generalized Linear Regression (GLM)) to act as benchmarks during the comparisons of the different learning algorithms under evaluation. Note that we have tested the majority of available standard ML algorithms in the *spark.ml* library in order to enrich the discussion of the results. Table 32 presents the hyperparameters considered for each regression model tested.

Table 32: Hyperparameter of the regression algorithms

Parameter	**Regression algorithms**					**Description**[a]
	DT	**RF**	**LR**	**GBT**	**GLM**	
maxDepth	✓	✓		✓		Maximum depth of the tree (*d=5*)
maxBins	✓	✓		✓		Maximum of bins for discretizing continuous features (*d=32*)
numTrees		✓				Number of trees to train (*d=20*)
elasticNetParam			✓			ElasticNet mixing parameter [0, 1](*d=0.0*)
regParam			✓		✓	Regularization parameter (*d=0.0*)
family					✓	description of error distribution (*d="gaussian"*)
maxIter			✓	✓	✓	Maximum number of interactions (*d=100*)

[a] Default Value (d)

5.4.5.1 Decision Tree (DT) regression

A Decision Tree (DT) is a commonly used algorithm for classification and regression tasks (Russel & Norving, 2010). It is a branching structure where several decision nodes are connected by branches that extend from the root node until the leaf nodes (Larose, 2005; Moro et al., 2014). Starting from the root node, attributes are tested at the decision nodes, resulting in a set of branches reflecting the possible outcomes. This branching representation can be translated into a set of IF-THEN statements, which shows its simplicity and therefore, easily understood by humans (Moro et al., 2014). In this work, we have adopted the Classification And Regression Tree (CART) algorithm (Loh, 2011) under variance reduction as split criterion (Sutton, 2005) to further compute information gain. In such a setting, for regression purposes, the final prediction results from the arithmetic average of the individuals trees. Despite their simplicity and ease of implementation, decision trees are extremely susceptible to overfitting, which motivates the use of more robust tree-based methods (e.g., random forest) based on bootstrap aggregation strategies able to minimize the prediction variance.

5.4.5.2 Random Forest (RF) regression

Random Forest (RF) is a popular and very efficient classification and regression algorithm, introduced by Breiman (2011) (Biau & Scornet, 2016; Couronné et al., 2018; Cutler et al., 2012; Genuer et al., 2010; Grömping, 2009). It consists of the aggregation of a large number of de-correlated decision trees built by the combination of bootstrap aggregation, consisting of random sampling with replacement to determine the individual tree estimates (Biau & Scornet, 2016), with random feature selection so that to reduce the variance, the risk of overfitting and achieve better prediction performance (Couronné et al., 2018; Spark, 2020; Witten et al., 2016). Each tree of RF is built using several bootstrap samples randomly chosen from the original dataset using the CART method and the Decrease Gini Impurity (DGI) as the splitting criterion. At each tree, the split is performed in such a way that the CART-criterion is maximized based on a given number of randomly selected candidate features (typically referred to as *mtry*)(Biau & Scornet, 2016; Couronné et al., 2018; Genuer et al., 2010).

5.4.5.3 Linear Regression (LR)

Linear Regression (LR) is a commonly used approach for modeling the relationship between scalar response, also known as the dependent variable, and one or more explanatory variables, also known as the independent variable. A linear regression is defined in the form:

$$Y = \beta_0 + X_1\beta_1 + ... + X_p\beta_p + \varepsilon = \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \varepsilon \tag{5.7}$$

where $\beta_0, \beta_1, ..., \beta_p$ are fixed and unknown parameters of the model (also known as regression coefficients), X_j represents the features (independent random variables), $j = 1, ..., p$ denote the regressor variables and ε denotes an error term (Gaussian random variable with expectation 0 and variance $\sigma^2 > 0$) (Grömping, 2009; Hastie et al., 2009). The most common method to estimate model coefficients is the Ordinary Least Squares (OLS) (Hastie et al., 2009). Yet, in this work, apart from the traditional OLS, we have considered L1 and L2 regularization techniques using the mean squared error as loss function. At this point, note that penalized regressions trade-off variance for bias, which has proven to be valuable in inventory management contexts (see, e.g., Kourentzes et al., 2020).

5.4.5.4 Generalized Linear Model (GLM)

Generalized Linear Model (GLM) is a flexible generalization of ordinary linear regression, formulated by Nelder and Wedderburn (1972). It generalizes the linear regression by permitting the linear model to be related to the response variable via a link function and permitting the magnitude of the variance of each measurement to be a function of its predicted value (Voyant et al., 2017). In this work, the main motivation for using GLM has to do with the nature of the LT variable. Since LT are discrete, GLM may reveal a better fit to non-Gaussian distributions when compared to that obtained by applying a linear regression (which assumes Gaussian distributions). In this context, when implementing the GLM algorithm, we tested the traditional Poisson and Gamma family distributions commonly adopted in fundamental inventory management textbooks (Silver et al., 2016) as well as the Tweedie distribution, which is able to not only approach some distributions in the exponential family but also accommodate null values (as is the case of LT for several suppliers geographically close to the case study company). In this process, we have excluded the Gaussian GLM as it is equivalent to the application of standard linear regression.

5.4.5.5 Gradient-Boosted Tree (GBT) regression

Gradient-Boosted Tree (GBT) is an ensemble method that combines a large set of decision trees to make a prediction (Ye et al., 2009). Boosting aims to iteratively combine several weak learners, aiming to generate a single strong learner. In this context, a weak learner is a tree whose performance is marginally higher than random chance. The addition of a new tree in each new iteration aims to correct the errors of the model built in the previous iteration. So, gradient boosting tries to fit a parameterized function (a tree) to pseudo-residuals, which are the gradient of the loss function being minimized over the training data, in order to construct additive regression models (J. H. Friedman, 2001). Incorporating randomization into the iterative procedure can improve the approximation accuracy and execution speed of gradient boosting. This randomness also increases the robustness against the overfitting (Burez & Van den Poel, 2009; J. Friedman, 2002; Hastie et al., 2009).

5.4.6 Feature preparation and hyperparameters tuning

5.4.6.1 Feature scaling

Feature scaling is an important issue to be tackled in the pre-processing stage in ML projects, mainly when working with several ML algorithms. Some ML algorithms are sensitive to feature scaling, while others are not. In this work, we adopted the Z-score standardization technique to transform each input variable in such a way to have zero mean and a unit standard deviation. When compared, for instance, to the classical range normalization (consisting of establishing new min-max limits for each variable), the Z-score transformation deals more effectively with outliers.

5.4.6.2 Feature selection

Feature selection aims at selecting a subset of features in order to remove all irrelevant and redundant features that may produce negative effects in the fitting process of the models as well as in the subsequent predictions. In this work, we adopted the Chi-Square test of independence (Greenwood & Nikulin, 1996), a nonparametric test that aims to assess whether there is an association between categorical variables. Nevertheless, it can be used for data with numerical inputs after a process of binning (or discretization), consisting in transforming numerical variables into categorical ones. For a given numerical variable, this is done by grouping its values into a small set of discrete values (bins), each representing a specific range of values of the original numerical variable. These bins can be calculated using, for instance, quartile-based thresholds. A limitation of this method, not in our particular case, is that it assumes a relatively large sample size.

5.4.6.3 Hyperparameters tuning

Hyperparameter tunning is an fundamental aspect when building effective machine learning models. Typically, there are two common approaches are used for this purpose, namely grid search and random search. While the former is a relatively slow approach, taking a lot of time searching for all possible combinations of parameters to find the optimal ones, the latter is a fast and effective one but makes a random search of parameters. In contrast, we have considered an alternative strategy based on Bayesian optimization using the HyperOpt Python library (Bergstra et al., 2015). Unlike grid search and random search, Bayesian optimization requires less iterations to generate the optimal combination of model hyperparameters. This strategy determines the next hyperparameter combination to be tested based on the previously-evaluated outcomes, reducing unnecessary evaluations and improving efficiency (see L. Yang & Shami, 2020, for details).

In order to make the hyperparameters optimization procedure more robust and less dependent on specific training-test windows, we combine the Bayesian optimization procedure with a timely-ordered 5-fold cross validation strategy (without shuffling). Following such setting, we test several tuples of hyperparameters in five ordered training and validation sets instead of using, for example, a single validation set obtained through an ordered holdout strategy. The entire optimization procedure appears summarized in Fig. 59. For each model, we define the hyperparameter space and set the Mean Absolute Error (MAE) metric as the criterion to be minimized over 10 iterations of the TPE method (Bergstra et al., 2011), which has yielded very interesting results in recent review studies (Shekhar et al., 2021). In the objective function, we employed a 5-fold cross-validation (as defined previously) to promote the selection of the best tuple of hyperparameters for each model using several train-validation configurations. The MAE metric for each model and hyperparameters configuration is evaluated over each validation set derived from the application of the 5-fold cross-validation.

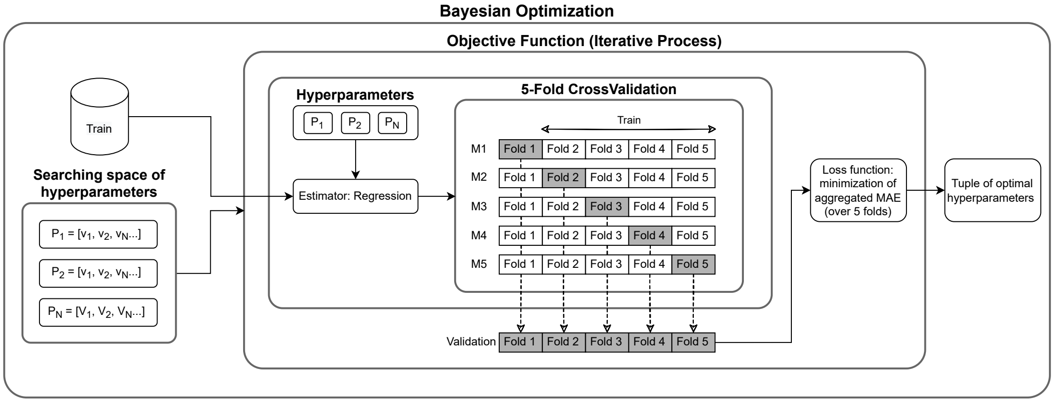

Bayesian Optimization
Objective Function (Iterative Process)
Train
Searching space of hyperparameters
P1 = [v1, v2, vN...]
P2 = [v1, v2, vN...]
PN = [V1, V2, VN...]
Hyperparameters
P1
P2
PN
Estimator: Regression
5-Fold CrossValidation
Train
M1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
M2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
M3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
M4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
M5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Validation Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Loss function: minimization of aggregated MAE (over 5 folds)
Tuple of optimal hyperparameters

Figure 59: Hyperparameter tuning using Bayesian Optimization.

5.4.7 Evaluation

To evaluate the regression models, we have adopted a realistic and robust RW scheme similar to that of Oliveira et al. (2017), as illustrated in Figure 60. In such a setting, we simulate a real environment by producing multiple training and testing iterations over time, rather than considering a simplistic holdout evaluation mechanism. This scheme works as follows. In this first iteration ($U = 1$), the model is trained using a window with a fixed size W to further generate T predictions. In the second iteration ($U = 2$), the training window of size W rolls forward S instances, causing the replacement of the S oldest instances of the training window by the S recent ones. The new model is retrained to further generate new subsequent T predictions. This process continues until the available samples is exhausted. The process of hyperparameters tunning takes place in each iteration of the RW scheme. Note that, as opposed to a typical rolling origin mechanism (Tashman, 2000), where the training window is incrementally increased, we discard the oldest observations as the training window rolls in time. Following this strategy, we avoid an excessive growth of the training window (containing now irrelevant vendor behavior dynamics) and thereby an increase in computational effort when training the models in each iteration of the evaluation process. The total of iterations is determined using the following closed-form expression:

$$U = (D - (W + T))/S. \tag{5.8}$$

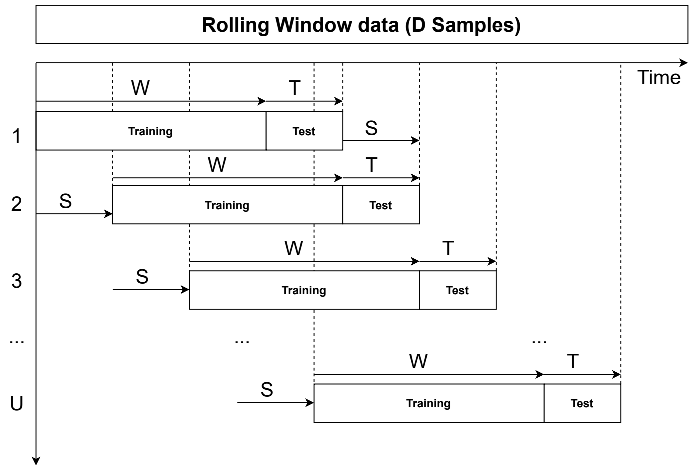


Figure 60: Rolling window scheme adapted from (Oliveira et al., 2017)

5.4.7.1 Measuring model performance

This section intends to report the two statistical metrics used to evaluate the performance of the different regression models under testing, namely the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). Such evaluation is conducted over the test sets of each rolling window iteration. The MAE and RMSE metrics are defined as follows:

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |Y_t - \widehat{Y_t}| \,, \tag{5.9}$$

where Y_t is the actual value and $\widehat{Y_t}$ is the target or predicted value.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t - \widehat{Y_t})^2} \,, \tag{5.10}$$

where Y_t is the actual LT and $\widehat{Y_t}$ is the predicted LT at t, whereas T is the total number of data points in each test set.

The MAE metric is one of the simplest and easily interpretable regression error metrics. It expresses the average of the absolute differences between the actual value and the predicted value. The MAE metric is also linear, which makes both larger and smaller errors contribute linearly to the overall error. On the other hand, RMSE metric punishes the larger deviations from the target value, thereby magnifying the overall error.

In addition to the above absolute and quadratic error metrics, we have also consider the Regression Error Characteristic (REC) curve, which summarizes the trade-off between the error tolerance (*x-axis*) and the percentage of points that are fit within the tolerance (*y-axis*), also known as the accuracy of a regression function (Bi & Bennett, 2003). The Area under the Regression Error Characteristic (AUREC) is calculated from the REC curve. The higher the AUREC, the better are the regression model estimations, while a perfect regression model has an AUREC of 1. The Area Over the REC curve (AOREC), which can be determined through the AUREC (AOREC = 1 - AUREC) represents the biased estimate of the expected error (Bi & Bennett, 2003).

5.4.7.2 Measuring model prediction bias

Prediction bias represents a measure of how far is the estimated value from the real value. In other words, a systematic deviation between estimated and real values. It is very useful to verify if a model tends to overestimate (where bias is less than zero) or underestimate (where bias is greater than zero) the predictions. In this work, following the interesting work of Barrow and Kourentzes (2016), we have determined the scaled Mean Errors (sME) and the scaled Squared Errors (sSE) of each model. Both sME, scaled Mean Squared Error ($sMSE$) are used as a complementary measure to the statistical metrics defined in Section 5.4.7.1.

$$sME = \sum_{t=1}^{T} \frac{Y_t - \widehat{Y}_t}{\overline{Y}} \tag{5.11}$$

$$sMSE = \sum_{t=1}^{T} \frac{(Y_t - \widehat{Y}_t)^2}{\overline{Y}} \tag{5.12}$$

where $\overline{Y}$ denotes the mean level of LT, Y_t is the real LT and $\widehat{Y}_t$ is the predicted value of LT at t.

5.4.7.3 Safety stock estimation

Safety stock also known as buffer stock represents an extra inventory that is held in order to deal with demand and supply uncertainties and avoid stock-outs (Barros et al., 2021). The safety stock of finished goods aims to attend unexpected demand and safety stock of raw material aims to protect against supply and production problems. In this work, we study the implications of both LT and demand variability on estimating the safety stocks of raw material. Among several traditional mathematical stochastic methods present in the literature studies, including supply chain books, we adopted a method that formulates the classical safety stock dimensioning problem, as follows:

$$SS = z\sqrt{\mu_{LT}\sigma_D^2 + (\mu_D\sigma_{LT})^2} \tag{5.13}$$

where $z = \Phi^{-1}[\alpha]$ represents the safety factor, $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function, μ_{LT}, σ_{LT} are the mean and standard deviation of LT, while μ_D, σ_D are the mean and variance of demand, respectively. This popular traditional method is very practical for determining the target safety stock, assuming that the mean and standard deviation of demand during lead time follow a Gaussian distribution. We are aware that the incorrect use of this assumption may result in higher inventory costs (see, Clark, 1957; Disney et al., 2016; Ruiz-Torres and Mahmoodi, 2010 for details). At this point, due to the lack of historical demand forecasting data from the case study company, we were not able to use the standard deviation of forecasting errors for safety stock dimensioning purposes.

To measure the inventory-related costs TC, in terms of the safety stocks estimated for each raw material m, we have adopted the following expression:

$$TC = (I_m \times P_m) \times SS_m \, , \tag{5.14}$$

where I_m is the holding rate for the raw material m per unit of time, P_m denotes the raw material standard unit price and SS_m represents the safety stock estimated.

5.5 Experiments and Results

5.5.1 Experimental and modeling setup

All experiments were performed using the framework Apache Spark 2.4.0, using code written in the PySpark (Python API of Apache Spark) and executed in the company Big Data cluster with a total of 9 nodes, 152 (144 with Hyper-threading) cores, 224 terabytes (TB) of disk capacity and 1603 gigabytes (GB) of memory capacity. For the computational experiments during the empirical evaluations, we have reserved 10GB of memory for the driver process, 7 executors with 4 cores to be used for each executor as well as 10GB to be used for each executor process. Moreover, as a *spark-submit* parameter, we set the deploy mode to cluster mode in order to execute the driver process inside the cluster (i.e., using the cluster resources).

We have explored five popular ML algorithms (RF, LR, GLM, GBT and DT) and also adopt the realistic and robust RW scheme for evaluating our models. After consulting logistics domain experts, we adopted the values *W = 21.000*, *T = 500* and *S = 500* for the RW scheme, producing a total of $U = 20$ iterations that allows the computation of aggregate results for each ML model using a statistical confidence measure. The modeling setup for each regression model is summarized as folows. For the DT model, we set *maxDepth* ∈

$\{2, 5, 10, 20, 30\}$ and *maxBins* $\in \{10, 20, 40, 80, 100\}$. In the case of RF, we additionally set the number of trees to *numTrees = 200*. The GBT was trained with 100 epochs (*maxIter = 100*) of L-BFGS algorithm (Liu & Nocedal, 1989) using the same hyperparameters as those established for DT. The LR was trained with 100 epochs (*maxIter = 100*) while considering *elasticNetParam* $\in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ and *regParam* $\in \{0.01, 0.1, 0.5, 1.0, 2.0\}$. This allows to account for regularized regressions (LASSO, Ridge Regression and Elastic Net). Finally, the GLM was also trained with 100 epochs considering *regParam* $\in \{0.01, 0.1, 0.5, 1.0, 2.0\}$, *family* $\in \{\text{``}poisson\text{''}, \text{``}gamma\text{''}, \text{``}tweedie\text{''}\}$ and the same parameters as DT to be optimized.

Finally, we take advantage of the non-parametric Wilcoxon signed-rank test (for paired samples) (Hollander et al., 2013) to compute a pseudo-median of the overall MAE (or RMSE) for each model over the RW iterations. We also check whether the prediction errors obtained over RW iterations are significantly different from one model to another. At this point, we give preference to the use of non-parametric tests over parametric ones as they make fewer assumptions about the underlying data.

5.5.2 Preliminary experiments

Before moving on to the application of the RW scheme to evaluate the overall performance of the regression models, we have used the data corresponding to the first iteration ($U = 1$) to perform a set of preliminary experiments for feature selection. Here, the goal is to assess the relevance of employing feature selection techniques in our particular machine learning pipeline. For that, we have employed an hold-out scheme on that first data portion, by dividing it in a timely-order fashion into training set (70%) and test set (30%). We considered the training set data to apply the Chi-Square test of independence (as defined in Section 5.4.6.2) aiming to select a fixed number of top features with high predictive power. We chose the number of top features as the criterion for the Chi-Square test, while considering several possibilities, namely *numTopFeatures* $\in \{81, 78, 74, 72, 69, 65, 60, 54, 46, 39, 30, 24\}$. The top features extracted from the application of the test are then considered as inputs to fit the different regression models and then evaluate their performance over the test set. For the sake of simplicity, we have considered the default hyperparameter values for all the regression models.

The results of the preliminary experiments are summarized in Fig. 61. On one hand, we can observe that the MAE tends to increase as we reduce the number of features used to fit the regression models. On the other hand, one can also notice that the overall levels of computational effort tend to remain relatively stable as we decrease the number of features considered. From these results, and for this particular dataset, we can conclude that the application of feature selection techniques does not bring significant benefits either in terms of predictive or computational performance. Hence, all the experiments hereinafter presented do not include feature selection. In other words, we have considered the full set of 21 features originally created (see Section 5.4.4.2). After the application of the one-hot encoding to the categorical variables, this set is increased to 81 features, which act as inputs for all regression models tested.

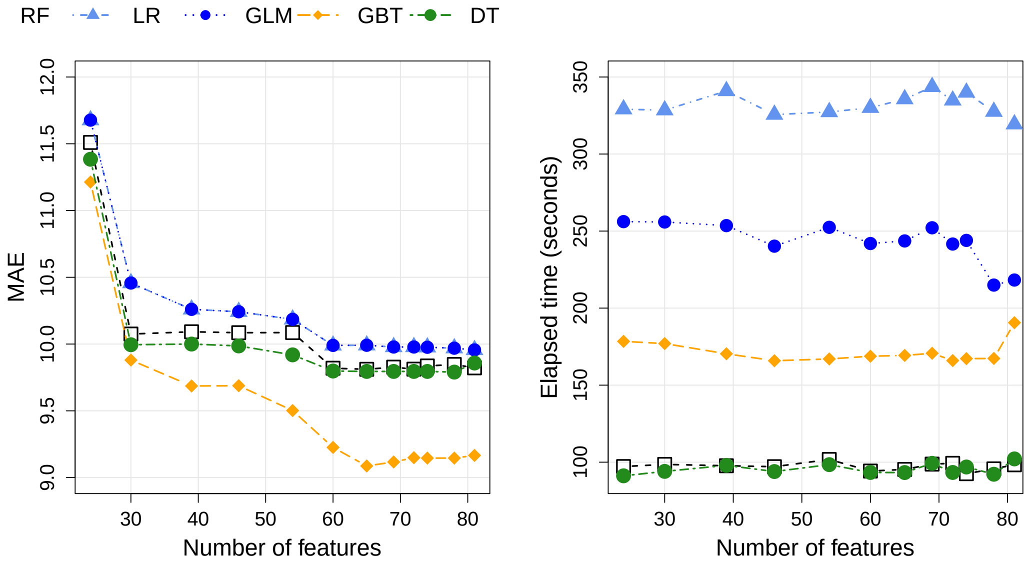


Figure 61: MAE and elapsed time (in seconds) as a function of the number of features included in the regression models.

5.5.3 Overall predictive performance

We start by summarizing the predictive performance of the tested regression models in terms of MAE and RMSE (Table 25). For the sake of model comparisons, we also evaluate the cost-accuracy trade-off by considering computation time metrics, namely training and prediction times. Following the RW schema presented in Section 4.2.5.6, each absolute/quadratic error recorded in Table 33 is the Wilcoxon pseudo-median (Hollander et al., 2013) of the different MAE (and RMSE) values in each RW iteration $U \in \{1, 2, ..., 20\}$.

Table 33: Comparison of the optimized ML models (best values in **bold**).

Models	MAE	RMSE	TT[a]	PT[b]
Random Forest (RF)	**7.378**[*]	**17.708**	871.087	14.582
Linear Regression (LR)	9.109	20.265	248.100	0.034
Generalized Linear Model (GLM)	9.026	20.310	184.335	0.036
Decision Tree (DT)	9.376	22.278	**71.426**	**0.028**
Gradient-Boosted Tree (GBT)	9.696	24.662	821.299	0.910

[*] Statistically significant under paired comparison with LR, GLM, DT and GBT

[a] TT: Training time (s)

[b] PT: Prediction time (s)

Focusing on the MAE values for simplicity, the results show that the RF model ranks first in terms of predictive capability, followed by GLM, LR, DT and GBT. However, it requires more time in both the training and prediction phases. In terms of overall forecasting accuracy, we further found that RF outperforms (with statistically significant differences under paired comparison, $p < 0.05$) all the remaining regression models. The evolution of the models predictive power, in terms of MAE, throughout the different iterations of the RW scheme is illustrated in Figure 62. One can observe that the RF model provides better predictive power at all RW iterations when compared with the remaining models. By way of example, Fig. 63 summarizes the good quality of the predictions at the RW iteration $U = 5$ of the RF model, which presents the plots of Coefficient of Determination (R^2) (left) and REC curve (right) computed with the corresponding AUREC curve. On the other hand, Figs. 64 and 65 depict the actual supply LT (measured) and the predicted supply LT (predicted) from the RF model for the RW iterations $U = 5$ and $U = 20$, respectively. By visual inspection, both figures show that the measured (actual) and predicted LT values are quite related in terms of magnitude. Interestingly, we observe that the RF generally tends to underestimate the predictions, especially for large values of supply LT. This behavior was motivated by the recent pandemic of Coronavirus (SARS-CoV-2), which caused dramatic increases in supplier delivery times. In these cases, the models naturally fail to capture such behaviors. In practice, underestimated supply LTs can lead to serious imbalances in inventory management and, in the limit, inventory stock-outs and service level damages to the customer. This result shows the importance of evaluating predictions not only in terms of conventional statistical error metrics but also in terms of prediction bias. Table 34 provides the overall prediction bias and the magnitude of prediction errors for the different models. At this point, while RF, LR and GLM tend to underestimate the predictions, the DT and GBT models tend to overestimate the predictions. Of note, when compared to LR and GLM models, RF underestimates the predictions in a low proportion. In light of these findings, we were interested in assessing which variables have the greatest impact on the predictive ability of the RF model. The ranking of the top 10 most important features for the RF model is depicted in Figure 66. Note that some of the features that most contribute to the model predictive power were created in the feature engineering process, stressing the added value of this modeling step in general predictive analytics tasks.

In order to facilitate the integration of the proposed approach in the case-study company, we developed a graphical interface in Power BI (Fig. 67). In this dashboard, besides showing the prediction of LT and the dimensioning of safety stocks for different service levels, it shows also a map with indication of the country where the order is dispatched and the destination country, several measures that support or contribute to the LT prediction (bottom boxes) and raw material information including supplier and forwarder information's (left boxes). The bottom boxes provide information regarding the median of supplier delays in a long-term fashion, the variation of LT over time and order volume over time for the selected raw material. Overall, this dashboard aims to support the decision making processes and improve the logistic planner's flexibility and productivity.

5.5.4 Safety stock performance

We have used Eq. (5.13) to estimate safety stock considering stochastic LT (using the RF model outcomes) and dynamic demand (using the manufacturer's demand obtained from MRP system). Fig. 68 depicts the required safety stock levels from different target α (Type I) service level of three different raw

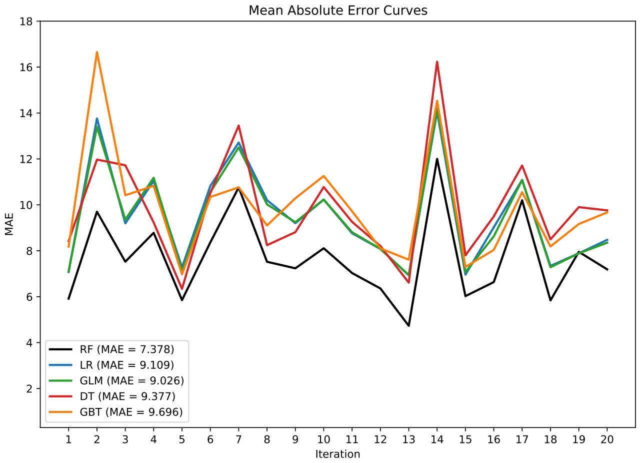


Figure 62: Evolution of MAE metric (*y*-axis) over the RW iterations for the different ML regression models (*x*-axis).

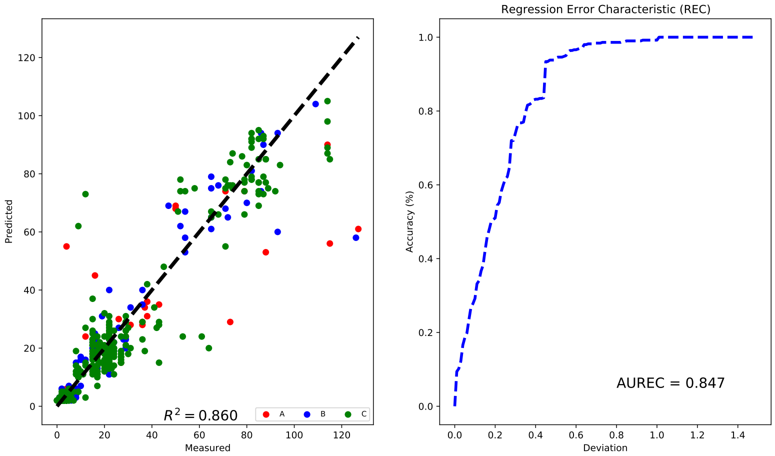


Figure 63: Coefficient of determination and REC curve at RW iteration *U*=*5* of the RF model.

Table 34: Models prediction bias (best values in **bold**).

Models	sME[a]	sMSE[b]
Random Forest (RF)	**0.076**	**17.798**
Linear Regression (LR)	0.115	25.645
Generalized Linear (GLM)	0.104	25.364
Decision Tree (DT)	-0.057	29.873
Gradient-Boosted Tree (GBT)	-0.049	29.751

[a] sME: scaled Mean Errors

[b] sMSE: scaled Mean Squared Errors

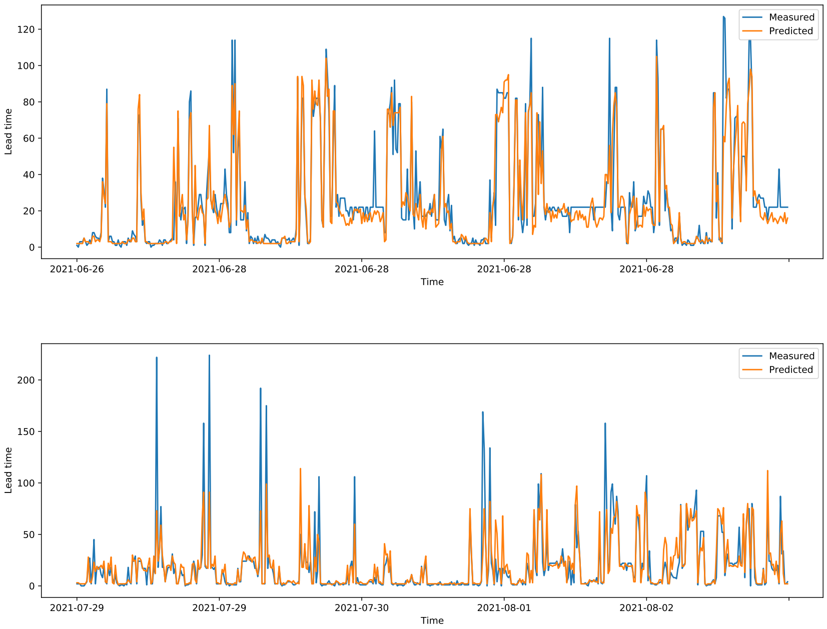


Figure 65: Measured vs Predicted Lead-time at RW iteration *U=20* of RF model.

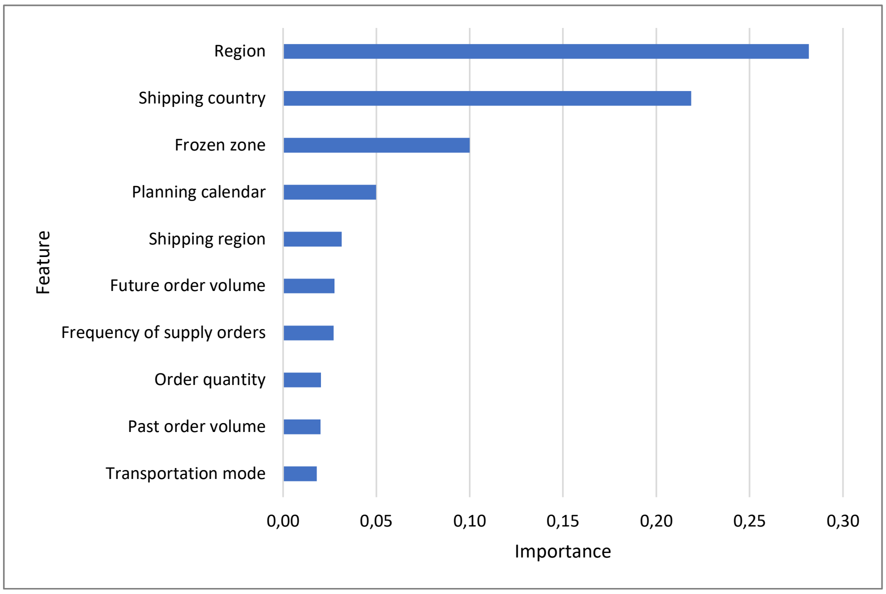
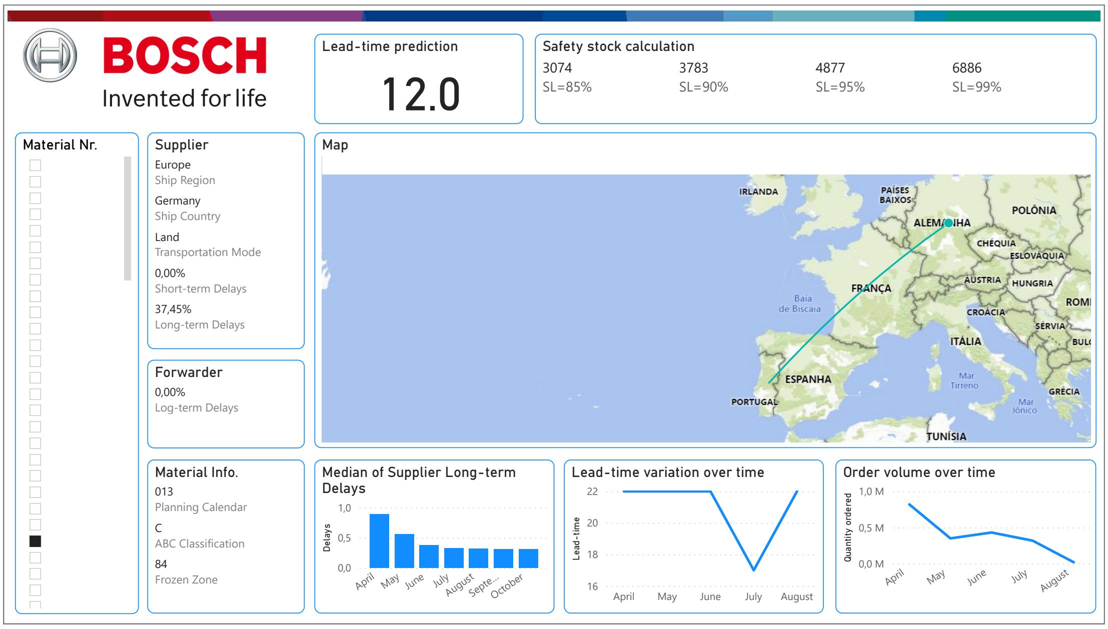


Figure 66: Feature importance for the RF model.

materials (R1, R2 and R3). We have tested the LT and demand for each of these raw materials using the Shapiro-Wilk test for normality (Yazici & Yolacan, 2007). The results have shown that, for all raw materials, there is no evidence to reject the hypothesis that both lead time and demand follow a Gaussian



BOSCH
Invented for life

Lead-time prediction
12.0

Safety stock calculation

3074	3783	4877	6886
SL=85%	SL=90%	SL=95%	SL=99%

Material Nr.

Supplier
Europe
Ship Region
Germany
Ship Country
Land
Transportation Mode
0,00%
Short-term Delays
37,45%
Long-term Delays

Forwarder
0,00%
Log-term Delays

Material Info.
013
Planning Calendar
C
ABC Classification
84
Frozen Zone

Figure 67: Dashboard: Lead-time prediction and safety stock estimation.

distribution.

As expected, Fig. 68, shows that safety stocks increase whenever service level increases. Table 35 describes the SS determined by the proposed approach and the current company approach. Here, as the case study company does not manage inventories and procurement processes according to a pre-specified service level, we only consider a single SS estimation from the case-study company. Comparing these two approaches, we can observe that, for raw material 3, the approach currently used by the company provides higher SS value than our calculation for different service levels. This is because the company prefers to set excessive levels of safety stock for class C materials, in light of their low holding cost. In contrast, for raw material 2, our calculation provides higher safety stocks than those derived from the actual approach used by the company, for all target service levels considered.

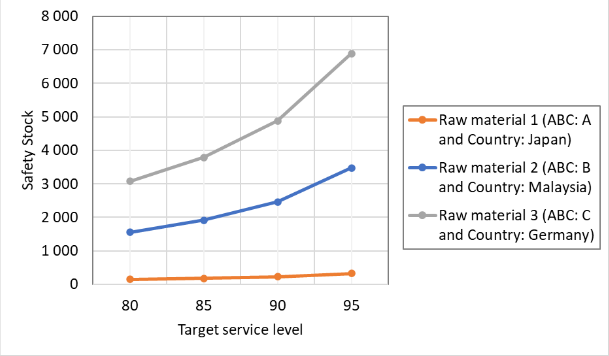


Figure 68: Safety stocks evolution as a function of target service levels.

Table 36 summarizes the safety stock costs derived from our calculations and from the the current

Table 35: Deviations in safety stocks (in units) with respect to the company estimation.

Raw Material	Proposed approach				Company estimation
	α-SL[d]= 85%	α-SL = 90%	α-SL = 95%	α-SL = 99%	
R1[a]	-49	-16	+37	+133	196
R2[b]	+173	+532	+1084	+2099	1380
R3[c]	-4328	-3618	-2524	-515	7401

[a]Raw material 1 (R1), [b]Raw material 2 (R2), [c]Raw material 3 (R3), Service Level (SL).

company approach. Regarding the raw material 1, we observed a reduction in safety stock holding costs of $\approx 25\%$ considering SL = 85% and $\approx 8\%$ considering α-SL = 90%. For the raw material 3, there is also a reduction of $\approx 58\%$ for α-SL = 85% and $\approx 49\%$ for α-SL = 90%. Contrarily, for the raw material R2, in agreement with the results presented in Table 35, there is a generalized increase of safety stock costs. In general, the company logistics managers tend to dimension the safety stock far beyond what is necessary, as a high customer service level is taken as a key indicator notwithstanding the associated inventory level. Hence, as long as customer service levels are not affected, a reduction in the safety stock levels translates into smaller inventory holding costs and, in this context, may allow for better inventory management by facilitating the process of material reception and its subsequent allocation to available warehouse areas.

Table 36: Deviations in safety stock costs with respect to the company estimation.

Raw Material	Proposed approach				Company approach
	α-SL[d]= 85%	α-SL = 90%	α-SL = 95%	α-SL = 99%	
R1[a]	-523	-161	397	3515	1423
R2[b]	2	6	11	21	13
R3[c]	-5	-4	-3	-1	8

[a] Raw material 1 (R1), [b] Raw material 2 (R2), [c] Raw material 3 (R3), Service Level (SL).

5.5.5 Practical & managerial implications

Our study shows the benefits of applying a multivariate supervised machine learning strategy to estimate supply lead time. An important aspect of our approach is that it is distribution-free and based on a big data framework, which facilitates its scalability and applicability within a real-world supply chain context. As lead time is a key parameter for dimensioning safety stocks, the fairly good performance of the proposed approach offers exciting opportunities to improve their calculation. However, since most of the studies focus on improving demand estimation processes, more studies are needed to model stochastic lead times and incorporate the uncertainty inherent to them in the calculation of dynamic safety stocks. We believe that this continues to be a research path of paramount importance in light of the current literature. Note that although the results produced are interesting in the sense of promoting a better modeling of supply lead time dynamics, the focus should also be given to initiatives that promote the improvement of demand estimations. Without acting in the demand forecasting domain, seeking to obtain stable forecasts, the bullwhip effect ratios tend to increase and, consequently, the development of approaches able to accommodate large supply order volatility becomes, in the limit, a challenging task.

When evaluation predictive models in a real-world context, we strongly support not just the use of conventional error metrics but also metrics that enable to measure the bias of the derived predictions. This plays a tremendous importance in inventory management contexts, allowing the decision maker to evaluate a predictive approach not only from the standpoint of its deviation from the actual value, but also from its tendency to overestimate or underestimate supply LT dynamics.

From a practical point of view, it should be emphasized that the application of machine learning approaches in real contexts entails some difficulties, namely regarding the existence of skilled human resources with the technical capabilities to successfully conduct and implement such approaches. This argument is particularly valid with respect to the implementation of machine learning models in a big data environment, for which there is not yet a sufficiently large set of parallelized strategies available for use by researchers and practitioners as those found in classical machine learning. On the other hand, as a final note, the practical implementation of the proposed approach requires that business experts take an active part in the process of searching for model improvements, especially when it comes to the inclusion/exclusion of variables that may or may not make sense in light of the supply chain dynamics at a given moment.

5.6 Summary

Uncertainty in lead time is a complex problem that affects the supply chain performance in terms of inventory levels and total costs (B. Dey et al., 2021). In this work, we propose a IDSS that combines machine learning and big data techniques to estimate supply LT and to estimate safety stocks using the derived estimations. The proposed IDSS was designed to use real-world data stored in a BDW. We build a machine learning pipeline that includes several steps. Firstly, regarding the data preparation step, we ingest the logistics data and construct our dataset, with a total of 32,000 records, to further conduct data cleansing tasks. Then, particular emphasis was given to feature engineering tasks in order to create new features from existing data using domain knowledge, with the aim of improving predictive capabilities. Secondly, within the feature preparation stage, the categorical features are converted to numerical ones and all features are standardized. Lastly, we evaluate five ML regression models: Random Forest, Decision Tree, Linear Regression, Generalized Linear Model and Gradient-Boosted regression Tree, adopting a realistic and robust RW scheme and using absolute and quadratic metrics to measure the overall performance of regression models. For our data, we found that the RF model provides the best predictive power, followed by the GLM, LR, DT and GBT. On the other hand, results from prediction bias have shown that, for our data, RF, LR and GLM tend to underestimate the LT prediction whereas DT and GBT tend to overestimate the predictions. Overall, RF model provides very interesting results both in terms of prediction bias and magnitude of predictions errors when compared to those obtained using the remaining regression models.

We also evaluate the inclusion of the derived supply LT estimations on a classical safety stock formulation. We evaluate the impacts of our approach by comparing it with the current approach adopted by the organization in terms of safety stock held and inventory holding costs. The results provide evidence regarding the usefulness of our approach in improving the safety stock estimation and minimizing the inventory holding costs. Finally, we developed a dashboard in the PowerBI tool to facilitate the use of the proposed decision support system by the Logistics planners, providing the estimations of supply LT and safety stock under different service levels.

From the practical point of view, the proposed decision support system provides several advantages, including the operational and financial performance of the case-study company. Moreover, it provides a systematic approach for estimating safety stock (instead of using a static and experience-based approach currently adopted by the case study company), proving to be valuable in supporting logistics decision-making process.

Promising research directions can be followed by coping with some limitations of our approach. Relevant research directions from this study are, for instance: (i) the incorporation of non-parametric demand forecasting approaches in safety stock estimations. Such approaches should be capable of capturing the real dynamics of SC demand over the product life-cycle; (ii) the adoption of empirical models that take advantage of dynamic lead times to further estimate dynamic safety stocks, rather than following our static SS proposals while assuming a parametric LT with mean μ_{LT} and variance σ^2_{LT}; (iii) the inclusion of a new set of relevant features for improving the predictive power of the proposed models; (iv) the inclusion of dummy variables to model supply LT peaks, as a way to improve their flexibility to highly nonlinear supply LT dynamics; (v) the testing of additional machine learning regression models (e.g., XGBoost and LightGBM) to our modeling experiments.

6 Conclusions

Summary: This last chapter concludes all work developed in this thesis. Firstly, the chapter initiates with the summary of all work developed. Then, a discussion is presented, including the description of the PhD limitations. Lastly, this chapter is closed with the presentation of future work directions.

Chapter Table of Contents:

6.1 Summary . . . 168
6.2 Discussion . . . 169
6.3 Future Work . . . 172

6.1 Summary

Over the last years, due to the introduction of the Industry 4.0 concept and the profound digital transformation of the industry, companies have expended significant time and effort to re-engineer their SC by changing their business process and technology focused on implementing integrated SCM. Logistics constitutes one of the crucial factors in the success of the SC. It consists of planning and coordinating the movement of products in a timely, safely and effectively way, and one of its main functions is to define suppliers for the raw materials and ensure on-time delivery of those raw materials at the right place and quantity, known as the Procurement process. The Procurement process as integral part of the Logistic process, directly impacts the organization's performance. Indeed, the problems associated with this process can result in high losses, such as losses of customer's confidence, reputation and revenues. The assessment of the Procurement problems and the creation of solutions to support this process allow the improvement of the efficiency of the supply chain and consequently improves the organizational performance. Thus, proper logistics management guides to improving the overall performance and, therefore, competitive advantage for organisations.

On the one hand, the digital transformation of industry associated with globalization and global market competitiveness contributes to transforming SC into an even more complex network and consequently more vulnerable to disruptions caused by several uncertainties and risks, including lead-time uncertainty, demand uncertainty, order cross-over, supplier delays, among others, which impacts the SC performance. Managing these uncertainties and risks is a fundamental challenge to organizations, and buffering techniques such as SS are naturally adopted by these organizations to comply with demand and supply LT uncertainties to protect against stock-outs. Therefore, supply LT constitutes a core parameter that affects the SC performance and represents an essential part of dimensioning a cost-effective safety stock. The importance of promoting better estimations of supply LT extends beyond the problem of estimating safety stocks so that it directly impacts the production planning and levelling, the management of capacities assigned to the production process, and the logistics and transport management. Nevertheless, delivery delays directly impact the supply LT and thus directly influence the overall inventory performance. Therefore, improving delivery performance is crucial for organizations.

On the other hand, the industry transformation into smart factories allows the generation of a massive amount of data, which can be analyzed in order to provide valuable insights for the organizations and also for optimising their SC. "Data is the oil"in terms of being one of the highly valuable resources for companies to improve the efficiency of their supply chain and consequently achieve organizational performance. In this context, BA is becoming increasingly essential value to industrial organizations, providing several benefits such as an increase in revenues, customer satisfaction and product quality, better resource planning, better insights on customer needs, optimized supply chain, better demand forecast, lower cost base (cost cutting), better compliance with regulations, among others (Božič & Dimovski, 2019; Lueth et al., 2016; Trkman et al., 2010).

The main contributions of this thesis consist of the two proposed approaches. Motivated by the identified gaps and opportunities reviewing the state-of-the-art related to dimensioning safety stocks under uncertainties and risks, we firstly drive our research on developing an ML-based framework for predicting the risk of supply delays in the Big Data context. This approach uses real-world empirical data to evaluate the model's predictive performance and the financial impact of model misclassication, contrary to common theoretical frameworks adopted. Naturally, supply delays represents the main factor of LT variability and strongly impact the inventory management (inventory stock-outs and damagees regarding the target service level) and demand management of the upstream SC side. Secondly, we focus on the main problem of interest, a novel IDSS for estimating supply lead time dynamics using a scalable technological Big Data architecture stood also proposed. It aims to improve the estimation of lead time uncertainty, to further promote better estimation of safety stock. This IDSS proven be valuable for logistic planners in optimizing safety stock levels and inventory levels comparaed to the current experience-based approach used by the case study company.

6.2 Discussion

The following provides a discussion of scientific contributions of this thesis, which can be divided into four categories.

A. **Systematic literature review on safety stock dimensioning under uncertainties and risks in the procurement process**

   In Chapter 3, a SLR was conducted regarding the topic of safety stock dimensioning allowing the identification of literature gaps and research opportunities to guide future research directions (see Section 3.6). One can conclude that safety stock-related problems remains an engaging topic for researchers and practitioners, and dimensioning of safety stock is the most addressed problem category among the remaining two categories (safety stock management and safety stock placement or allocation or positioning). Moreover, overall, optimization is the most used technique to tackle this problem.

   Regarding the gaps in the literature and the research opportunities identified, it is worth highlighting the need to model supply LT variability and then take that into account when dimensioning safety stocks. Motivated by the lack of research studies in this field, we investigate a novel IDSS for predicting supply LT dynamics and ultimately promote better safety stock estimations.

B. **Supervised learning approach for predicting supply delay risks**

   Chapter 4 proposed a novel ML-based approach for predicting the risk of supply delay using a scalable technological BDA architecture. It focuses on identifying the risk of supply delay in a proactive manner by combining several variables whose dynamics may affect positively or negatively the supplier performance, in detriment to the risk response manner vastly proposed in the literature. A ML pipeline stood proposed and therefore used as a basis to carry out all ML experiments in the BD environment. Contrarily to most research studies that have overlooked supply chain performance metrics, this one proposed was evaluated in terms of predictive power (statistical accuracy) and misclassification-related costs. Hence, the selection of the ML model was based not only on the criterion of providing better predictive power but also considering the minimum inventory-related costs caused by the mistaken classifications of the model. Likewise, it is noteworthy to highlight the importance of business domain expert inputs regarding several tasks of a data mining project, mainly in the feature engineering process, which was crucial for the current predictive capacity of the models. Overall, the proposed approach offers several benefits, mainly in operational and financial performance for the organization. It proves to be very useful for logistics planners, helping them in decision-making process and enabling proactive actions regarding possible supply delays. Moreover, this approach can help in improving the efficiency of the inventory management process through cost savings and ensuring proper control of logistics performance. However, some limitations are pointed out, such as the size of the dataset, creation of new features and automatic mechanisms to deploy and retrain ML models, lack of transparency of black-box ML methods, and the applicability to other industries.

C. **Supervised learning approach for estimating supply lead times**

In Chapter 5, a multivariate supervised approach to estimate supply LT is proposed using ML and BD techniques. Motivated by the existing gap in the literature, data-driven approaches to model supply LT were explored, regarding the enhanciment the logistics decision-making process. The proposed approach incorporates several variables that potentially impacts supply LT regarding a given supplier over time. Promoting better estimations of supply LT is crucial for better dimensioning of safety stock, but also highly impacts the production planning and leveling, management of capacities allocated to the production process, and logistics and transport management. Similarly to the approach in Chapter 4, a ML pipeline is also proposed to guide the ML tasks. Nevertheless, the importance of assessing the predictive ability of ML models not solely by conventional statistical error metrics, but also in measuring the prediction bias, was highlighted. This measurement is crucial in the context of inventory management, enabling logistics planners to visualise the deviation of predicted values from the actual and the tendency to over- or under-estimate supply LT dynamics. Thus, underestimating supply LT can lead to serious imbalances in inventory management and, even inventory stock-outs and damage in customer service level. On the other hand, overestimating supply LT can lead to an excess inventory and therefore extra inventory holding costs. The proposed approach has proven to be very useful for logistic planners, helping them in decision-making process. The importance of business experts in the practical implementation of the proposed approach is highlighted, as they actively participate in the process of seeking model improvements, especially with regard to both feature selection and engineering tasks. The main limitations corroborate with the need for a skilled human resource with the technical ability to conduct and implement such approaches in a big data environment and not take into account the crossover phenomenon when modeling supply LT.

D. **Intelligent Decision Support System towards improved safety stock dimensioning**
In Chapter 5, a IDSS is proposed also to estimate safety stocks using the derived estimations. It combines a machine learning for providing LT predictions (described above) and big data techniques to use real-world data stored in a BDW. Moreover, the proposed IDSS provides a systematic approach to determine safety stock under dynamic manufacturer demand and stochastic LT, minimizing the holding costs while attending to a certain service level. Currently, the case study company use a static and experience-based approach to estimate safety stock. Figure 69 summarizes the research flow up to the IDSS towards safety stock estimations. Firstly, it starts with a given safety stock dimensioning-related problem, a research gap that was previously highlighted (*1. Literature Analysis*). In effect, our research is motivated by the difficulties of determining upstream (manufacturer's) variations of supply lead-time towards safety stock dimensioning improvement. Then, following the Business understanding, the business domain expert knowledge is fundamental, helping to understand the issue under consideration and its main impact factors. A BDW to promote the Logistics 4.0 movement and thus improve the organization's performance were proposed and implemented. It aims to store real-world logistics-related data from Bosch AE/P, mainly data of purchases, orders, order deliveries, and product inventory, in order to allow to be used as a basis to address both lead-time and supply delay problems. Note that the development of the BDW is framed within the research project in which this thesis is as well inserted, and both of approaches proposed in Chapter 4 and 5 are used to validated the BDW. In Chapter 4, supplier delay risk is addressed as the main factor impacting supply LT. Delivery delays have a direct impact on the overall inventory management performance. Thus, dealing with it is crucial to be improve the level of customer service and, consequently, the organization's performance. Data quality analysis was performed to check data cardinality, correlation, missing values and zeros, followed by the exploratory data analysis to explore the on-hand dataset. Then, the domain expert's knowledge was crucial for creating new variables from the original data that positively or negatively affect the supplier's delivery performance. Afterwards, feature preparation is performed and several ML models were evaluated in terms of statistical accuracy through the rolling window scheme. However, the selection of the best ML model is determinated by considering the statistical accuracy and the impact of misclassification error, as illustrated in the 6. (the inventory performance validation method is shown by using the dashed box). The features that contribute most to the prediction of supply delay are used as input to the proposed approach in Chapter 5. In summary, regarding the supervised learning approach in Chapter 5, the ML models are explored and thus evaluated using quadratic metrics, as well as the prediction bias regarding LT predictions. However, before this, tasks such as dataset creation, data quality and data exploration were carried out. Furthermore,

model selection was performed based on models performance evaluations using regression metrics, and also prediction bias and magnitude of predictions errors. Finally, the predicted LT values from the selected regression model were used to estimate safety stocks. Our research derives dynamic estimations of SS by assuming stochastic LT, in contrast to the vast majority of research studies in the literature that assumes LT to be deterministic. The impact of the proposed IDSS was measured in terms of SS estimations and invetory holding costs, and therefore compared with the current experience-based approach adopted by Bosch AE/P. This experience-based approach is grounded in experience and is quite prone to errors, besides requiring a lot of knowledge of the business process. Overall, it should be emphasized that the proposed IDSS proves to be valuable to the organization, supporting in the decision-making process of logistics planners. However, some of the limitations of our approach relies in the incorporation of a non-parametric demand forecasting approaches in the safety stock estimations, adoption of empirical non-parametric models for SS estimations, and considering crossover phenomenon in modeling LT.

6.3 Future Work

This section is devoted to future work that can be followed regarding this Doctor of Philosophy (PhD) thesis. The objectives initially defined are successfully achieved (see Section 1.2), but there is still a place for further explorations and contributions, as follows identified:

- **Applicability** - This thesis was conducted at Bosch AE/P to address issues faced by the organization. Thereby, two IT artifacts were procuded as described in Section 4 and 5. For both artifacts considered proprietary data from the case study company. However, their applicability can be extrapolated to other several areas of activity, such as manufacturing and electronics components (semicondutor) industries, in order to compare with the results obtained from our experiments.

- **Automated data extraction and models deployment** - In future work, automated deployment of ML models proposed in Chapter 4 and 5 should be considered in order to decrease the human intervention. As such, Apache Oozie could be used as the worflow coordenator to schedule and runs the workflow of Apache Hadoop jobs (see Section 5.4.2). Therefore, the Oozie workflow should be parameterized to be triggered based on regular time intervals, data availability and/or external events, in order to allow automated training, testing, refining and re-training of the proposed ML models in the company BD cluster. Furthermore, is also important to consider a larger data set with more historical data collected by the organization. Likewise, is necessary to improve and automate the data extraction on a daily basis. External data (e.g., weather data or natural and human-made disasters) could also be considered for addressing supply chain risks and uncertainties.

- **Feature engineering** - Feature engineering is a core part of any ML task, which seeks to generate a set of new features on the basis of raw data, to enhance the prediction powers of the ML models

Domingos (2012). The importance of this task is also supported by the benefits it brought to the models developed in Chapter 4 and 5, since many of the new features created were among the features that contributed most to the predictions. Future research should focus on creating new features to improve the ML predictions. *"(...) features that look irrelevant in isolation may be relevant in combination. (...) there is ultimately no replacement for the smarts you put into feature engineering"* (Domingos, 2012). Moreover, automated feature engineering could be also be explored to complement feature engineering performed in manual manner.

- **Explainable machine learning** - The supervised black-box ML methods, such as SVM, ANN, Deep Learning (DL), RF, eXtreme Gradient-Boosting Machine (XGBoost), among others, are increasingly being used to address several problem regarding different areas of activity, allowing powerful and accurate predictions (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020). These models are too complex and lack transparency, i.e., can not be directly explained or easily understood by a human. Nevertheless, in general, humans are sceptical about adopting techniques that are not directly interpretable, tractable and trustworthy. This situation worsens when an important decision must be entrusted to a system that can not explain the basis of its decisions (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020). Yet, Explainable Artificial Intelligence (XAI) proposes to address these issues by making AI more transparent. Future research could focus on improving the interpretability of these ML approaches by exploring XAI approaches, using libraries such as Local Interpretable Model-Agnostic Explanations (Lime)[1] and SHapley Additive exPlanations (SHAP)[2].

- **Machine learning algorithms** - In Chapter 4 and 5, several standard ML algorithms available in the *spark.ml* library were considered for classification and regression tasks. However, it would also be interesting to test other ML algorithms not natively available in Apache Spark for establishing benchmarks, such as XGBoost and Light Gradient Boosting Machine (LightGBM).

- **Automated machine learning** - The development of typical ML models always requires a human resource with skills to perform all the essential tasks of ML, which include data pre-processing, feature engineering, feature extraction, feature selection, algorithm selection and hyperparameter optimisation tasks (Feurer et al., 2015). However, applying each of these tasks appropriately in a manual manner may be challenging and a very time-consuming process. Thus, the first mechanisms for automating machine learning emerged, which the objective to produce test set predictions automatically without human effort, and within a fixed computational budget (Feurer et al., 2015). Given the lack of studies regarding the application of Automated Machine Learning (AutoML)-based approaches in inventory management-related problems, it could be considered and exploited for subsequent comparison with the results of the proposed methods (typical development of ML models). AutoML tools such as H2O AutoML[3] could be explored for benchmarking purposes.

[1] https://lime-ml.readthedocs.io/en/latest/

[2] https://shap.readthedocs.io/

[3] https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

- **Estimating safety stock** - Non-parametric empirical models that take advantage of dynamic LT and demand forecast errors to further estimate dynamic safety stocks should be employed instead of the static closed-form formulations that assume a parametric lead time with mean μ_{LT} and variance σ^2_{LT} and demand with mean μ^2_D and variance σ^2_D (see Section 5.4.7.3). However, current non-parametric models estimate safety stock under the past lead time demand forecast errors.

- **Demand forecasting** - Several inventory control models proposed in the literature, including fundamental inventory management textbooks, assume that the demand distribution is known and deterministic (i.e., known in advance) (Prak & Teunter, 2019). In Chapter 5, we also consider manufacturer's demand as deterministic for safety stock determination. However, it is well known that this assumption is not realistic in real-world supply chain environments. Thus, in practice, demand-related information is not available, and future demand must be predicted based on historical observations (Prak & Teunter, 2019; P. Silva et al., 2022). Incorporating a non-parametric approach to estimate real demand dynamics into safety stock estimates represents a fascinating subject for future research. Furthermore, considering PLC and seasonality in demand variation remains a good and challenging topic. Current technological advances have made PLC ever shorter and therefore more difficult to estimate, especially in areas of activity such as the mobile phone, fashion (shoes and clothing) and electronics components industries. Therefore, accurate demand forecasting is crucial for real-world applications, especially regarding inventory management, and in particular, to estimate safety stocks (Barrow & Kourentzes, 2016).

- **Lead time with order crossover** - It is common sense that LT characteristics represent the main parameter for estimating safety stock, and thus, its distortions can have a significant impact on the supply chain performance measures. Several literature studies have been devoted to stochastic lead time modelling for decades (see, for example, Scarf (1958), Silver et al. (2016), and Zipkin (2000)), nevertheless, only a few have considered the stochastic LT with the presence of order crossover phenomenon (e.g., Chatfield and Pritchard (2018), Riezebos (2006), and Srinivasan et al. (2011)). Generally, little attention has been devoted to it and often this aspect is even neglected. The order crossover phenomenon is increasingly prone to occur, especially in the modern supply chains, much because of its exposition to several uncertainties and risks. The stochastic LT problems are very difficult to investigate, and and associated with the order crossover phenomenon makes it especially difficult. Subsequently, it would be interesting to consider this phenomenon in the approach proposed in Chapter 5 for estimating LT.

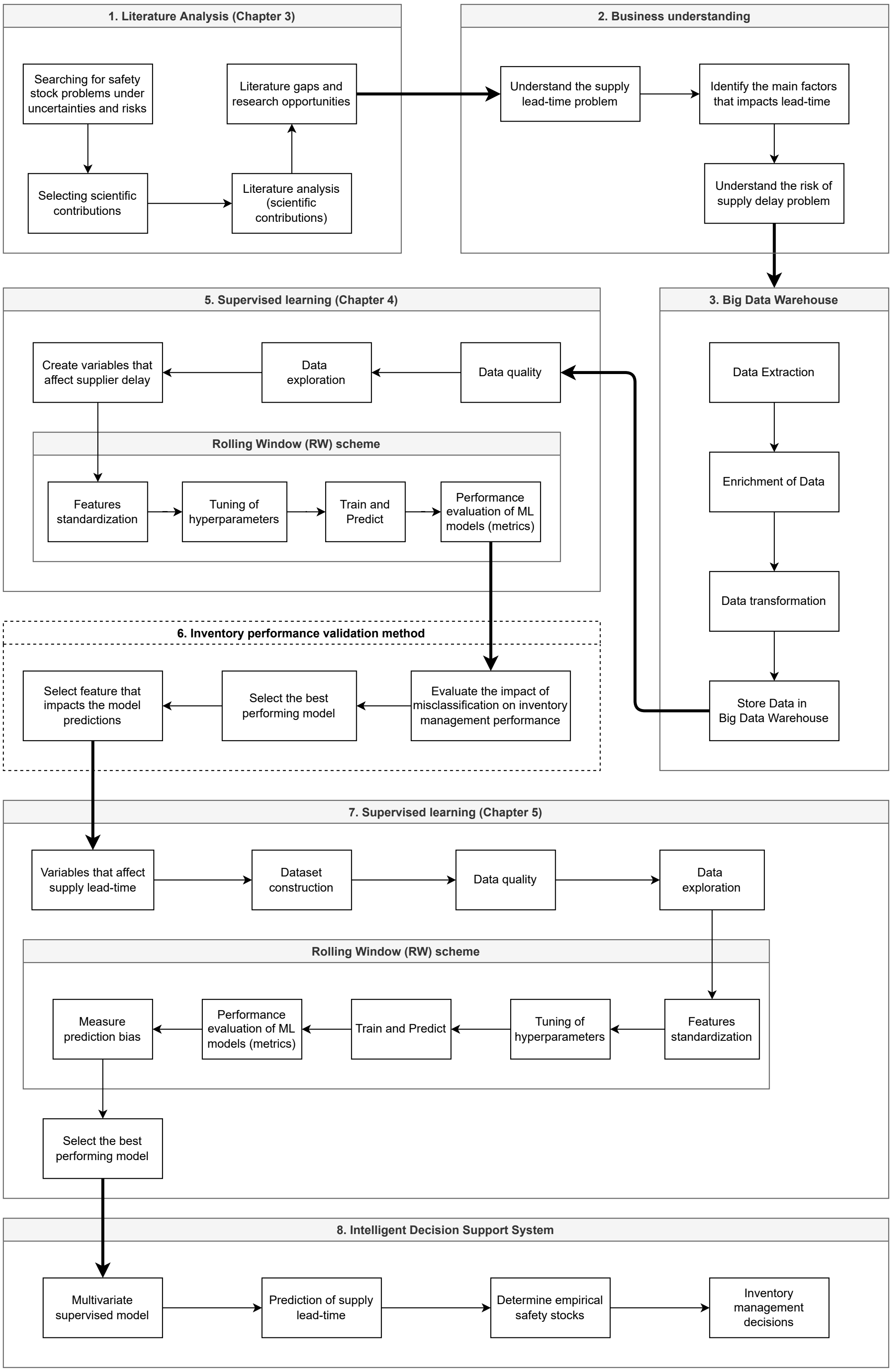

1. Literature Analysis (Chapter 3)
Searching for safety stock problems under uncertainties and risks
Selecting scientific contributions
Literature analysis (scientific contributions)
Literature gaps and research opportunities
2. Business understanding
Understand the supply lead-time problem
Identify the main factors that impacts lead-time
Understand the risk of supply delay problem
3. Big Data Warehouse
Data Extraction
Enrichment of Data
Data transformation
Store Data in Big Data Warehouse
5. Supervised learning (Chapter 4)
Data quality
Data exploration
Create variables that affect supplier delay
Rolling Window (RW) scheme
Features standardization
Tuning of hyperparameters
Train and Predict
Performance evaluation of ML models (metrics)
6. Inventory performance validation method
Evaluate the impact of misclassification on inventory management performance
Select the best performing model
Select feature that impacts the model predictions
7. Supervised learning (Chapter 5)
Variables that affect supply lead-time
Dataset construction
Data quality
Data exploration
Rolling Window (RW) scheme
Features standardization
Tuning of hyperparameters
Train and Predict
Performance evaluation of ML models (metrics)
Measure prediction bias
Select the best performing model
8. Intelligent Decision Support System
Multivariate supervised model
Prediction of supply lead-time
Determine empirical safety stocks
Inventory management decisions

Figure 69: Intelligent Decision Support System for safety stock dimensioning.

Bibliography

Abdel-Malek, L., Kullpattaranirun, T., & Nanthavanij, S. (2005). A framework for comparing outsourcing strategies in multi-layered supply chains. *International Journal of Production Economics*, *97*(3), 318–328. https://doi.org/10.1016/j.ijpe.2004.09.001 (cit. on pp. 2, 57, 59, 138).

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 (cit. on p. 173).

Adenso-Diaz, B. (1996). How many units will be short when stockout occurs? *International Journal of Operations & Production Management*, *16*(4), 112–&. https://doi.org/10.1108/01443579610114121 (cit. on p. 58).

Aftab, U., & Siddiqui, G. (2018). Big data augmentation with data warehouse: A survey. *2018 IEEE International Conference on Big Data (Big Data)*, 2775–2784. https://doi.org/10.1109/BigData.2018.8622182 (cit. on p. 84).

Albrecht, M. (2014). Determining near optimal base-stock levels in two-stage general inventory systems. *European Journal of Operational Research*, *232*(2), 342–349. https://doi.org/10.1016/j.ejor.2013.07.025 (cit. on pp. 69, 70).

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Communications Surveys and Tutorials*, *17*(4), 2347–2376. https://doi.org/10.1109/COMST.2015.2444095 (cit. on pp. 16, 17).

Alguliyev, R., Imamverdiyev, Y., & Sukhostat, L. (2018). Cyber-physical systems and their security issues. *Computers in Industry*, *100*(April), 212–223. https://doi.org/10.1016/j.compind.2018.04.017 (cit. on p. 17).

Alicke, K. (2005). *Planung und Betrieb von Logistiknetzwerken* (2nd ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/3-540-27748-X (cit. on p. 44).

Altendorfer, K. (2019). Effect of limited capacity on optimal planning parameters for a multi-item production system with setup times and advance demand information. *International Journal of Production Research*, *57*(6), 1892–1913. https://doi.org/10.1080/00207543.2018.1511925 (cit. on pp. 57, 58).

Angkiriwang, R., Pujawan, I. N., & Santosa, B. (2014). Managing uncertainty through supply chain flexibility: reactive vs. proactive approaches. *Production and Manufacturing Research*, *2*(1), 50–70. https://doi.org/10.1080/21693277.2014.882804 (cit. on pp. 42, 43).

Aravinth, S., Begam, A. H., Shanmugapriyaa, S., Sowmya, S., & Arun, E. (2015). An efficient hadoop frameworks sqoop and ambari for big data processing. *International Journal for Innovative Research in Science and Technology*, *1*(10), 252–255 (cit. on pp. 88, 109).

Armbrust, M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K., Meng, X., Kaftan, T., Franklin, M. J., Ghodsi, A., & Zaharia, M. (2015). Spark sql: Relational data processing in spark. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 1383–1394. https://doi.org/10.1145/2723372.2742797 (cit. on pp. 89, 109).

Arnott, D., & Pervan, G. (2016). A critical analysis of decision support systems research revisited: The rise of design science. *Enacting Research Methods in Information Systems: Volume 3*, *29*(4), 43–103. https://doi.org/10.1007/978-3-319-29272-4_3 (cit. on pp. xiii, 18–20).

Avci, M., & Selim, H. (2017). A Multi-objective, simulation-based optimization framework for supply chains with premium freights. *Expert Systems with Applications*, *67*, 95–106. https://doi.org/10.1016/j.eswa.2016.09.034 (cit. on pp. 60, 61, 67, 122).

Avci, M., & Selim, H. (2018). A multi-objective simulation-based optimization approach for inventory replenishment problem with premium freights in convergent supply chains. *Omega (United Kingdom)*, *80*, 153–165. https://doi.org/10.1016/j.omega.2017.08.016 (cit. on pp. 66, 67).

Axsäter, S. (2015). *Inventory Control* (Vol. 225). Springer International Publishing. https://doi.org/10.1007/978-3-319-15729-0 (cit. on pp. 46, 75).

Babai, M., Syntetos, A., Dallery, Y., & Nikolopoulos, K. (2009). Dynamic re-order point inventory control with lead-time uncertainty: analysis and empirical investigation. *International Journal of Production Research*, *47*(9), 2461–2483. https://doi.org/10.1080/00207540701666824 (cit. on pp. 65, 66).

Bahroun, Z., & Belgacem, N. (2019). Determination of dynamic safety stocks for cyclic production schedules. *Operations Management Research*, *12*(1-2), 62–93. https://doi.org/10.1007/s12063-019-00140-0 (cit. on pp. 57, 58).

Baily, P., Farmer, D., Crocker, B., Jessop, D., & Jones, D. (2015). *Procurement: Principles and Management* (5th ed.). Edition, Eleventh. (Cit. on pp. 43, 104).

Balfaqih, H., Nopiah, Z. M., Saibani, N., & Al-Nory, M. (2016). Review of supply chain performance measurement systems: 1998–2015. *Computers in Industry*, *82*, 135–150. https://doi.org/10.1016/j.compind.2016.07.002 (cit. on pp. 2, 37).

Balster, A., Hansen, O., Friedrich, H., & Ludwig, A. (2020). An ETA Prediction Model for Intermodal Transport Networks Based on Machine Learning. *Business and Information Systems Engineering*, *62*(5), 403–416. https://doi.org/10.1007/s12599-020-00653-0 (cit. on pp. 106, 108).

Bandaly, D., Satir, A., & Shanker, L. (2016). Impact of lead time variability in supply chain risk management. *International Journal of Production Economics*, *180*, 88–100. https://doi.org/10.1016/j.ijpe.2016.07.014 (cit. on p. 137).

Baranowski, Z., Grzybek, M., Canali, L., Garcia, D. L., & Surdy, K. (2015). Scale out databases for cern use cases. *J. Phys. Conf. Ser*, *664*(4), 042–002 (cit. on pp. 88, 109, 144).

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012 (cit. on p. 173).

Barros, J., Cortez, P., & Carvalho, M. S. (2021). A systematic literature review about dimensioning safety stock under uncertainties and risks in the procurement process. *Operations Research Perspectives*, *8*(January), 100192. https://doi.org/10.1016/j.orp.2021.100192 (cit. on pp. 138, 156).

Barrow, D., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, *177*, 24–33. https://doi.org/10.1016/j.ijpe.2016.03.017 (cit. on pp. 156, 174).

Baryannis, G., Dani, S., & Antoniou, G. (2019). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems*, *101*, 993–1004. https://doi.org/10.1016/j.future.2019.07.059 (cit. on pp. 41, 103–108).

Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: State of the art and future research directions. *International Journal of Production Research*, *57*, 2179–2202. https://doi.org/10.1080/00207543.2018.1530476 (cit. on pp. 103, 107).

Ben-Ammar, O., Bettayeb, B., & Dolgui, A. (2019). Optimization of multi-period supply planning under stochastic lead times and a dynamic demand. *International Journal of Production Economics*, *218*, 106–117. https://doi.org/10.1016/j.ijpe.2019.05.003 (cit. on pp. 61, 62).

Benbitour, M. H., Sahin, E., & Dallery, Y. (2019). The use of rush deliveries in periodic review assemble-to-order systems. *International Journal of Production Research*, *57*(13), 4078–4097. https://doi.org/10.1080/00207543.2018.1505059 (cit. on pp. 55, 58).

Bender, J., & Ovtcharova, J. (2021). Prototyping machine-learning-supported lead time prediction using automl. *Procedia Computer Science*, *180*, 649–655. https://doi.org/10.1016/j.procs.2021.01.287 (cit. on p. 141).

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *30th International Conference on Machine Learning, ICML 2013*, (PART 1), 115–123 (cit. on p. 119).

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, 1–9 (cit. on pp. 119, 154).

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., & Cox, D. (2015). Hyperopt: A python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, *8*(1), 014008 (cit. on p. 153).

Berling, P., & Marklund, J. (2014). Multi-echelon inventory control: An adjusted normal demand model for implementation in practice. *International Journal of Production Research*, *52*(11), 3331–3347. https://doi.org/10.1080/00207543.2013.873555 (cit. on p. 63).

Beutel, A., & Minner, S. (2012). Safety stock planning under causal demand forecasting. *International Journal of Production Economics*, *140*(2), 637–645. https://doi.org/10.1016/j.ijpe.2011.04.017 (cit. on pp. 56, 58).

Bi, J., & Bennett, K. (2003). Regression error characteristic curves. *Proceeding of the Twentieth International Conference on Machine Learning (ICML-2003)* (cit. on p. 156).

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197–227. https://doi.org/10.1007/s11749-016-0481-7 (cit. on pp. 151, 152).

Bittorf, M., Bobrovytsky, T., Erickson, C., Hecht, M. G. D., Kuff, M., Leblang, D. K. A., Robinson, N., Rus, D. R. S., Wanderman, J., & Yoder, M. M. (2015). Impala: A modern, open-source sql engine for hadoop. *Proceedings of the 7th biennial conference on innovative data systems research* (cit. on pp. 88, 89, 109, 110).

Bollapragada, R., Rao, U., & Zhang, J. (2004). Managing inventory and supply performance in assembly systems with random supply capacity and demand. *Management Science*, *50*(12), 1729–1743. https://doi.org/10.1287/mnsc.1040.0314 (cit. on p. 73).

Bossert, J., & Willems, S. (2007). A periodic-review modeling approach for guaranteed service supply chains. *Interfaces*, *37*(5), 420–435. https://doi.org/10.1287/inte.1070.0298 (cit. on pp. 69, 70).

Boulaksil, Y. (2016). Safety stock placement in supply chains with demand forecast updates. *Operations Research Perspectives*, *3*, 27–31. https://doi.org/10.1016/j.orp.2016.07.001 (cit. on p. 71).

Boulaksil, Y., Fransoo, J., & Van Halm, E. (2009). Setting safety stocks in multi-stage inventory systems under rolling horizon mathematical programming models. *OR Spectrum*, *31*(1), 121–140. https://doi.org/10.1007/s00291-007-0086-3 (cit. on pp. 55, 58).

Boyes, H., Hallaq, B., Cunningham, J., & Watson, T. (2018). The industrial internet of things (IIoT): An analysis framework. *Computers in Industry*, *101*(June), 1–12. https://doi.org/https://doi.org/10.1016/j.compind.2018.04.015 (cit. on pp. 15, 16, 18).

Božič, K., & Dimovski, V. (2019). Business intelligence and analytics for value creation: The role of absorptive capacity. *International Journal of Information Management*, *46*(February 2018), 93–103. https://doi.org/10.1016/j.ijinfomgt.2018.11.020 (cit. on pp. 22, 23, 103, 168).

Bradley, J., & Robinson, L. (2005). Improved Base-Stock Approximations for Independent Stochastic Lead Times with Order Crossover. *Manufacturing & Service Operations Management*, *7*(4), 319–329. https://doi.org/10.1287/msom.1050.0085 (cit. on p. 43).

Braglia, M., Castellano, D., & Frosolini, M. (2014). Safety stock management in single vendor-single buyer problem under VMI with consignment stock agreement. *International Journal of Production Economics*, *154*, 16–31. https://doi.org/10.1016/j.ijpe.2014.04.007 (cit. on pp. 66, 67).

Braglia, M., Castellano, D., & Frosolini, M. (2016). A novel approach to safety stock management in a coordinated supply chain with controllable lead time using present value. *Applied Stochastic Models in Business and Industry*, *32*(1), 99–112. https://doi.org/10.1002/asmb.2126 (cit. on pp. 44, 63, 64).

Braglia, M., Gabbrielli, R., & Zammori, F. (2013). Stock diffusion theory: A dynamic model for inventory control. *International Journal of Production Research*, *51*(10), 3018–3036. https://doi.org/10.1080/00207543.2012.752584 (cit. on pp. 63, 64).

Brander, P., & Forsberg, R. (2006). Determination of safety stocks for cyclic schedules with stochastic demands. *International Journal of Production Economics*, *104*(2), 271–295. https://doi.org/10.1016/j.ijpe.2004.11.009 (cit. on pp. 55, 58).

Breiman, L. (2011). Random forests. *Machine Learning*, (45), 5–32. https://doi.org/10.1023/A:1010933404324 (cit. on pp. 117, 126, 151).

Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P., & McFarlane, D. (2020). Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing. *International Journal of Production Research*, *58*(11), 3330–3341. https://doi.org/10.1080/00207543.2019.1685705 (cit. on pp. 103, 104, 106–108, 114, 148, 149).

Brunaud, B., Laínez-Aguirre, J., Pinto, J., & Grossmann, I. (2019). Inventory policies and safety stock optimization for supply chain planning. *AIChE Journal*, *65*(1), 99–112. https://doi.org/10.1002/aic.16421 (cit. on p. 63).

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3, Part 1), 4626–4636. https://doi.org/https://doi.org/10.1016/j.eswa.2008.05.027 (cit. on p. 153).

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167 (cit. on p. 117).

Burggräf, P., Wagner, J., Koke, B., & Steinberg, F. (2020). Approaches for the prediction of lead times in an engineer to order environment-a systematic review. *IEEE Access*, *8*, 142434–142445. https://doi.org/10.1109/ACCESS.2020.3010050 (cit. on p. 140).

Buzacott, J. (1999). Dynamic inventory targets revisited. *Journal of the Operational Research Society*, *50*(7), 697–703. https://doi.org/10.1057/palgrave.jors.2600744 (cit. on pp. 63, 64).

Caceres, H., Yu, D., & Nikolaev, A. (2018). Evaluating shortfall distributions in periodic inventory systems with stochastic endogenous demands and lead-times. *Annals of Operations Research*, *271*(2), 405–427. https://doi.org/10.1007/s10479-018-2764-8 (cit. on pp. 66, 67).

Campbell, G. (1995). Establishing safety stocks for master production schedules. *Production Planning and Control*, *6*(5), 404–412. https://doi.org/10.1080/09537289508930297 (cit. on pp. 60, 61).

Cao, D., & Silver, E. (2005). A dynamic allocation heuristic for centralized safety stock. *Naval Research Logistics*, *52*(6), 513–526. https://doi.org/10.1002/nav.20093 (cit. on pp. 70, 71).

Caridi, M., & Cigolini, R. (2002a). Improving materials management effectiveness: A step towards agile enterprise. *International Journal of Physical Distribution and Logistics Management*, *32*(7), 556–576. https://doi.org/10.1108/09600030210442586 (cit. on pp. 38, 44, 48, 54, 62, 67).

Caridi, M., & Cigolini, R. (2002b). Managing safety and strategic stocks to improve materials requirements planning performance. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *216*(7), 1061–1065. https://doi.org/10.1243/09544050260174256 (cit. on pp. 55, 58).

Carvalho, J. C., Guedes, A. P., Arantes, A. J. M., Martins, A. L., Póvoa, A. P. B., Luís, C. A., Dias, E. B., Dias, J. C. Q., de Menezes, J. C. R., Ferreira, L. M. D. F., Carvalho, M. d. S., Oliveira, R. C., Azevedo, S. G., & Ramos, T. (2017). *Logística e Gestão da Cadeia de Abastecimento* (2nd ed.). Edições Sílabo, Lda. (Cit. on pp. xiii, xiv, 2, 27–30, 37, 39–42, 62).

Cavalcante, I., Frazzon, E., Forcellini, F., & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing [cited By 60]. *International Journal of Information Management*, *49*, 86–97. https://doi.org/10.1016/j.ijinfomgt.2019.03.004 (cit. on pp. 106, 108).

Chaharsooghi, S. K., & Heydari, J. (2010). LT variance or LT mean reduction in supply chain management: Which one has a higher impact on SC performance? *International Journal of Production Economics*, *124*(2), 475–481. https://doi.org/10.1016/j.ijpe.2009.12.010 (cit. on pp. 2, 137, 140).

Chan, G. (1997). Eyeballing heuristics for dynamic lot sizing problems with rolling horizons. *Computers and Operations Research*, *24*(4), 379–385. https://doi.org/10.1016/S0305-0548(96)00039-1 (cit. on p. 58).

Chandra, C., & Grabis, J. (2008). Inventory management with variable lead-time dependent procurement cost. *Omega*, *36*(5), 877–887. https://doi.org/10.1016/j.omega.2006.04.009 (cit. on p. 65).

Chang, C. A. (1985). The interchangeability of safety stocks and safety lead time. *Journal of Operations Management*, *6*(1), 35–42. https://doi.org/10.1016/0272-6963(85)90033-6 (cit. on pp. 38, 43, 137).

Chang, W., & Lin, Y. (2019). The effect of lead-time on supply chain resilience performance. *Asia Pacific Management Review*, *24*, 298–309. https://doi.org/10.1016/j.apmrv.2018.10.004 (cit. on p. 137).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM -Cross-Industry Standard Process for Data Mining- 1.0 Step-by-step data mining guide. *CRISP-DM Consortium*, 76. https://doi.org/10.1109/ICETET.2008.239 (cit. on pp. xiii, xiv, 7–9, 105, 106).

Chatfield, D., Kim, J., Harrison, T., & Hayya, J. (2004). The bullwhip effect-impact of stochastic lead time, information quality, and information sharing: A simulation study. *Production and Operations Management* (cit. on p. 137).

Chatfield, D., & Pritchard, A. (2018). Crossover aware base stock decisions for service-driven systems. *Transportation Research Part E - Logistics and Transportation Review*, *114*, 312–330. https://doi.org/10.1016/j.tre.2017.12.013 (cit. on pp. 43, 66, 67, 75, 174).

Chaturvedi, A., & Martínez-De-Albéniz, V. (2016). Safety Stock, Excess Capacity or Diversification: Trade-Offs under Supply and Demand Uncertainty. *Production and Operations Management*, *25*(1), 77–95. https://doi.org/10.1111/poms.12406 (cit. on pp. 61, 62).

Chen, H., & Li, P. (2015). Optimization of (R, Q) policies for serial inventory systems using the guaranteed service approach. *Computers and Industrial Engineering*, *80*, 261–273. https://doi.org/10.1016/j.cie.2014.12.003 (cit. on pp. 69, 70).

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly: : Management Information Systems*, *36*(4), 1165–1188 (cit. on p. 19).

Chen, Y., Pekny, J., & Reklaitis, G. (2013). Integrated planning and optimization of clinical trial supply chain system with risk pooling. *Industrial and Engineering Chemistry Research*, *52*(1), 152–165. https://doi.org/10.1021/ie300823b (cit. on pp. 55, 58).

Chevalier, M., Malki, M. E., Kopliku, A., Teste, O., & Tournier, R. (2015). Implementing multidimensional data warehouses into nosql. *17th International Conference on Enterprise Information Systems (ICEIS 2015) held in conjunction with ENASE 2015 and GISTAM 2015*, 172–183. http://publications.ut-capitole.fr/29466/ (cit. on p. 85).

Choi, T. M., Govindan, K., Li, X., & Li, Y. (2017). Innovative supply chain optimization models with multiple uncertainty factors. *Annals of Operations Research*, *257*(1-2), 1–14. https://doi.org/10.1007/s10479-017-2582-4 (cit. on p. 42).

Chopra, S., & Meindl, P. (2016). *Supply Chain Management: Strategy, Planning, and Operation*. Person Education Limited. (Cit. on pp. 2, 26–29, 37, 39, 42, 45, 46, 75, 103, 136).

Chopra, S., Reinhardt, G., & Dada, M. (2004). The effect of lead time uncertainty on safety stocks. *Decision Sciences*, *35* (cit. on pp. 2, 138).

Chou, S., Yang, C., Jiang, F., & Chang, C. (2018). The implementation of a data-accessing platform built from big data warehouse of electric loads. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, *02*, 87–92. https://doi.org/10.1109/COMPSAC.2018.10208 (cit. on p. 85).

Christopher, M. (2003). *Understanding Supply Chain Risk : A Self-Assessment Workbook*. Cranfield University, Cranfield School of Management, Centre for Logistics; Supply Chain Management. (Cit. on p. 41).

Chu, L., & Shen, Z. (2010). A power-of-two ordering policy for one-warehouse multiretailer systems with stochastic demand. *Operations Research*, *58*(2), 492–502. https://doi.org/10.1287/opre.1090.0707 (cit. on pp. 63, 64).

Chung, S., Kang, H., & Pearn, W. (2005). A service level model for the control wafers safety inventory problem. *International Journal of Advanced Manufacturing Technology*, *26*(5-6), 591–597. https://doi.org/10.1007/s00170-003-2028-9 (cit. on p. 61).

Chung, W., Talluri, S., & Kovács, G. (2018). Investigating the effects of lead-time uncertainties and safety stocks on logistical performance in a border-crossing jit supply chain. *Computers and Industrial Engineering*, *118*, 440–450. https://doi.org/10.1016/j.cie.2018.03.018 (cit. on p. 138).

Clark, C. (1957). Mathematical Analysis of an Inventory Case. *Operations Research*, *5*(5), 627–643. https://doi.org/10.1287/opre.5.5.627 (cit. on pp. 44, 157).

Cobb, B. (2016). Inventory control for returnable transport items in a closed-loop supply chain. *Transportation Research Part E: Logistics and Transportation Review*, *86*, 53–68. https://doi.org/10.1016/j.tre.2015.12.010 (cit. on p. 65).

Coleman, B. (2000). Determining the correct service level target [cited By 10]. *Production and Inventory Management Journal*, *41*(1), 19–23 (cit. on pp. 45, 46).

Colicchia, C., & Strozzi, F. (2012). Supply chain risk management: A new methodology for a systematic literature review. *Supply Chain Management*, *17*(4), 403–418. https://doi.org/10.1108/13598541211246558 (cit. on p. 40).

Colombo, A., Karnouskos, S., Kaynak, O., Shi, Y., & Yin, S. (2017). Industrial Cyberphysical Systems: A Backbone of the Fourth Industrial Revolution. *IEEE Industrial Electronics Magazine*, *11*(1), 6–16. https://doi.org/10.1109/MIE.2017.2648857 (cit. on pp. 17, 18).

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6171 LNAI*, 572–583. https://doi.org/10.1007/978-3-642-14400-4_44 (cit. on p. 117).

Costa, C., Andrade, C., & Santos, M. Y. (2018). Big Data Warehouses for Smart Industries. In *Encyclopedia of big data technologies* (pp. 1–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-63962-8_204-1 (cit. on pp. 88, 109).

Costa, C., & Santos, M. Y. (2017a). Big Data: State-of-the-art concepts, techniques, technologies, modeling approaches and research challenges. *IAENG International Journal of Computer Science*, *44*(3), 285–301 (cit. on p. 103).

Costa, C., & Santos, M. Y. (2017b). The suscity big data warehousing approach for smart cities. *Proceedings of the 21st International Database Engineering; Applications Symposium*, 264–273. https://doi.org/10.1145/3105831.3105841 (cit. on p. 86).

Costa, C., & Santos, M. Y. (2018). Evaluating several design patterns and trends in big data warehousing systems. *International Conference on Advanced Information Systems Engineering*, 459–473 (cit. on p. 85).

Council of Supply Chain Management Professional. (2010). *Supply Chain Operations Reference (SCOR) model* (tech. rep.). https://doi.org/10.15358/9783800639960_559 (cit. on p. 24).

Council of Supply Chain Management Professional - CSCMP. (2013a). CSCMP Supply Chain Management Definitions and Glossary. Retrieved 2019-03-12, from https://cscmp.org/CSCMP/Educate/SCM_Definitions_and_Glossary_of_Terms/CSCMP/Educate/SCM_Definitions_and_Glossary_of_Terms.aspx (cit. on pp. 23, 24, 26, 37, 39).

Council of Supply Chain Management Professional - CSCMP. (2013b). Supply Chain Management Concepts. Retrieved 2019-03-28, from https://cscmp.org/CSCMP/Develop/Starting_Your_SCM_Career/SCM_Concepts/CSCMP/Develop/Starting_Your_Career/Supply_Chain_Management_Concepts.aspx?hkey=96af0d8b-21ad-4bca-b7d1-956a25ced524&fbclid=IwAR1Yr3qvbmxdg8QxxgltXqFkxXOr6-7-ccQusCMJMNnsrSCaG7cH6InVMaA (cit. on pp. xiii, 24, 25).

Council of Supply Chain Management Professional - CSCMP. (2013c). Supply Chain Management Concepts. Retrieved 2019-03-28, from https://cscmp.org/CSCMP/Develop/Starting%7B%5C_%7DYour%7B%5C_%7DSCM%7B%5C_%7DCareer/SCM%7B%5C_%7DConcepts/CSCMP/Develop/Starting%7B%5C_%7DYour%7B%5C_%7DCareer/Supply%7B%5C_%7DChain%7B%5C_%7DManagement%7B%5C_%7DConcepts.aspx?hkey=96af0d8b-21ad-4bca-b7d1-956a25ced524%7B%5C&%7Dfbclid=IwAR1Yr3qvbmxdg8QxxgltXqFkxXOr6-7-ccQusCMJMNnsrSCaG7cH6InVMaA (cit. on p. 136).

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*. https://doi.org/10.1186/s12859-018-2264-5 (cit. on pp. 151, 152).

Cutler, A., Cutler, D., & Stevens, J. (2012). Random forests. Springer US. https://doi.org/10.1007/978-1-4419-9326-7 (cit. on p. 151).

De Smet, N., Aghezzaf, E.-H., & Desmet, B. (2019). Optimising installation (R,Q) policies in distribution networks with stochastic lead times: a comparative analysis of guaranteed- and stochastic service models. *International Journal of Production Research*, *57*(13), 4148–4165. https://doi.org/10.1080/00207543.2018.1518606 (cit. on pp. 72, 73).

Dean, J., & Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, *51*(1), 107–113. https://doi.org/10.1145/1327452.1327492 (cit. on pp. 88, 109).

de Armas, J., & Laguna, M. (2019). Parallel machine, capacitated lot-sizing and scheduling for the pipe-insulation industry. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2019.1600763 (cit. on p. 60).

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, *55*(1), 359–363. https://doi.org/10.1016/j.dss.2012.05.044 (cit. on pp. 19, 20).

Desmet, B., Aghezzaf, E., & Vanmaele, H. (2010). A normal approximation model for safety stock optimization in a two-echelon distribution system. *Journal of the Operational Research Society*, *61*(1), 156–163. https://doi.org/10.1057/jors.2008.150 (cit. on p. 73).

Dey, B., Bhuniya, S., & Sarkar, B. (2021). Involvement of controllable lead time and variable demand for a smart manufacturing system under a supply chain management. *Expert Systems with Applications*, *184*. https://doi.org/10.1016/j.eswa.2021.115464 (cit. on pp. 2, 137, 165).

Dey, O. (2019). A fuzzy random integrated inventory model with imperfect production under optimal vendor investment. *Operational Research*, *19*(1), 101–115. https://doi.org/10.1007/s12351-016-0286-1 (cit. on pp. 56, 58).

Digiesi, S., Mossa, G., & Mummolo, G. (2013). Supply lead time uncertainty in a sustainable order quantity inventory model. *Management and Production Engineering Review*, *4*, 15–27. https://doi.org/10.2478/mper-2013-0034 (cit. on pp. 2, 57, 59, 138).

Disney, S., Farasyn, I., Lambrecht, M., Towill, D., & de Velde, W. (2006). Taming the bullwhip effect whilst watching customer service in a single supply chain echelon. *European Journal of Operational Research*, *173*(1), 151–172. https://doi.org/10.1016/j.ejor.2005.01.026 (cit. on p. 42).

Disney, S., Maltz, A., Wang, X., & Warburton, R. D. (2016). Inventory management for stochastic lead times with order crossovers. *European Journal of Operational Research*, *248*(2), 473–486. https://doi.org/10.1016/j.ejor.2015.07.047 (cit. on pp. 2, 44, 66, 67, 138, 157).

Dolgui, A., & Prodhon, C. (2007). Supply planning under uncertainties in mrp environments: A state of the art. *Annual Reviews in Control*, *31*, 269–279. https://doi.org/10.1016/j.arcontrol.2007.02.007 (cit. on pp. 2, 137).

Domingos, P. (2012). A few useful things to know about machine learning. https://doi.org/10.1145/2347736.2347755 (cit. on pp. 20, 114, 115, 148, 173).

Doreswamy, Gad, I., & Manjunatha, B. R. (2017). Hybrid data warehouse model for climate big data analysis. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, 1–9. https://doi.org/10.1109/ICCPCT.2017.8074229 (cit. on p. 86).

Drath, R., & Horch, A. (2014). Industrie 4.0: Hit or hype? [Industry Forum]. *IEEE Industrial Electronics Magazine*, *8*(2), 56–58. https://doi.org/10.1109/MIE.2014.2312079 (cit. on pp. xiii, 15, 16, 22, 103).

Duc, T., Luong, H., & Kim, Y. (2008). A measure of the bullwhip effect in supply chains with stochastic lead time. *International Journal of Advanced Manufacturing Technology*, *38*, 1201–1212. https://doi.org/10.1007/s00170-007-1170-1 (cit. on p. 137).

Duong, A., Vo, V., Carvalho, M., Sampaio, P., & Truong, H. (2022). Risks and supply chain performance: Globalization and covid-19 perspectives. *International Journal of Productivity and Performance Management*, *ahead-of-print*. https://doi.org/10.1108/IJPPM-03-2021-0179 (cit. on p. 137).

Dwivedi, A. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, *29*, 685–693 (cit. on p. 118).

Elkan, C. (2001). The foundations of cost-sensitive learning. *IJCAI International Joint Conference on Artificial Intelligence*, 973–978 (cit. on p. 121).

Epstein, R., Neely, A., Weintraub, A., Valenzuela, F., Hurtado, S., Gonzalez, G., Beiza, A., Naveas, M., Infante, F., Alarcon, F., Angulo, G., Berner, C., Catalan, J., Gonzalez, C., & Yung, D. (2012). A strategic empty container logistics optimization in a major shipping company. *Interfaces*, *42*(1), 5–16. https://doi.org/10.1287/inte.1110.0611 (cit. on p. 73).

Er Kara, M., Oktay Fırat, S., & Ghadge, A. (2020). A data mining-based framework for supply chain risk management [cited By 18]. *Computers and Industrial Engineering*, *139*. https://doi.org/10.1016/j.cie.2018.12.017 (cit. on pp. 103–105, 137).

Eruguz, A. S., Sahin, E., Jemai, Z., & Dallery, Y. (2016). A comprehensive survey of guaranteed-service models for multi-echelon inventory optimization. *International Journal of Production Economics*, *172*, 110–125. https://doi.org/10.1016/j.ijpe.2015.11.017 (cit. on p. 38).

Fahimnia, B., Tang, C., Davarzani, H., & Sarkis, J. (2015). Quantitative models for managing supply chain risks: A review. *European Journal of Operational Research*, *247*(1), 1–15. https://doi.org/10.1016/j.ejor.2015.04.034 (cit. on p. 103).

Fahimnia, B., Tang, C. S., Davarzani, H., & Sarkis, J. (2015). Quantitative models for managing supply chain risks: A review. *European Journal of Operational Research*, *247*(1), 1–15. https://doi.org/10.1016/j.ejor.2015.04.034 (cit. on p. 48).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010 (cit. on pp. 116, 120, 121).

Feng, K., Rao, U., & Raturi, A. (2011). Setting planned orders in master production scheduling under demand uncertainty. *International Journal of Production Research*, *49*(13), 4007–4025. https://doi.org/10.1080/00207543.2010.495955 (cit. on p. 58).

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. Efficient and robust automated machine learning. In: *2015-January*. 2015, 2962–2970 (cit. on p. 173).

Fichtinger, J., Chan, C., & Yates, N. (2019). A joint network design and multi-echelon inventory optimisation approach for supply chain segmentation. *International Journal of Production Economics*, *209*, 103–111. https://doi.org/10.1016/j.ijpe.2017.09.003 (cit. on p. 71).

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*, 367–378. www.elsevier.com/locate/csda (cit. on p. 153).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1–39 (cit. on p. 153).

Funaki, K. (2012). Strategic safety stock placement in supply chain design with due-date based demand. *International Journal of Production Economics*, *135*(1), 4–13. https://doi.org/10.1016/j.ijpe.2010.11.015 (cit. on pp. 68–71).

Galar, M., Fern, A., Barrenechea, E., & Bustince, H. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(4), 463–484 (cit. on p. 120).

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007 (cit. on p. 21).

Gansterer, M., Almeder, C., & Hartl, R. (2014). Simulation-based optimization methods for setting production planning parameters. *International Journal of Production Economics*, *151*, 206–213. https://doi.org/10.1016/j.ijpe.2013.10.016 (cit. on p. 58).

Genuer, R., Poggi, J., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*, 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014 (cit. on pp. 151, 152).

Ghadge, A., Kara, M. E., Moradlou, H., & Goswami, M. (2020). The impact of industry 4.0 implementation on supply chains. *Journal of Manufacturing Technology Management* (cit. on pp. 82, 83).

Ghafour, K. (2018). Optimising safety stocks and reorder points when the demand and the lead-time are probabilistic in cement manufacturing. *International Journal of Procurement Management*, *11*(3), 387–398. https://doi.org/10.1504/IJPM.2018.091672 (cit. on p. 73).

Ghiani, G., Laporte, G., & Musmanno, R. (2004). *Introduction to Logistics Systems Planning and Control* (Advisors Editors, Ed.; 1st ed.). John Wiley & Sons. (Cit. on pp. xiii, 28, 29).

Gibson, B. J., Defee, C. C., Hanna, J. B., & Chen, H. (2014). *The definitive guide to Integrated Supply Chain Management: Optimize the interaction between Supply Chain Processes, Tools, and Technologies* (Council of Supply Chain Management Professionals, Ed.; 1st ed.). Pearson Education. https://www.pearson.com/us/higher-education/program/CSCMP-Definitive-Guide-to-Integrated-Supply-Chain-Management-The-Paperback/PGM2399248.html (cit. on p. 25).

Glock, C. (2012). Lead time reduction strategies in a single-vendor-single-buyer integrated inventory model with lot size-dependent lead times and stochastic demand. *International Journal of Production Economics*, *136*(1), 37–44. https://doi.org/10.1016/j.ijpe.2011.09.007 (cit. on pp. 56, 58).

Goldsby, T., Iyengar, D., & Rao, S. (2014). *The definitive guide to transportation* (Council of Supply Chain Management Professionals, Ed.; 1st ed.). Pearson Education. (Cit. on p. 29).

Gonçalves, J., Carvalho, M. S., & Cortez, P. (2020). Operations research models and methods for safety stock determination: A review [cited By 1]. *Operations Research Perspectives*, *7*. https://doi.org/10.1016/j.orp.2020.100164 (cit. on pp. 132, 138).

Gonçalves, J., Cortez, P., Carvalho, M., & Frazão, N. (2021). A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain [cited By 0]. *Decision Support Systems*, *142*. https://doi.org/10.1016/j.dss.2020.113452 (cit. on p. 75).

Govindan, K., Cheng, T., Mishra, N., & Shukla, N. (2018). Big data analytics and application for logistics and supply chain management. *Transportation Research Part E: Logistics and Transportation Review*, *114*(March), 343–349. https://doi.org/10.1016/j.tre.2018.03.011 (cit. on pp. 22, 103).

Grace Hua, N., & Willems, S. (2016). Analytical insights into two-stage serial line supply chain safety stock. *International Journal of Production Economics*, *181*, 107–112. https://doi.org/10.1016/j.ijpe.2015.10.010 (cit. on p. 71).

Grahl, J., Minner, S., & Dittmar, D. (2016). Meta-heuristics for placing strategic safety stock in multi-echelon inventory with differentiated service times. *Annals of Operations Research*, *242*(2), 489–504. https://doi.org/10.1007/s10479-014-1635-1 (cit. on pp. 69, 70).

Graves, S., & Schoenmeyr, T. (2016). Strategic safety-stock placement in supply chains with capacity constraints. *Manufacturing and Service Operations Management*, *18*(3), 445–460. https://doi.org/10.1287/msom.2016.0577 (cit. on pp. 72, 73).

Graves, S., & Willems, S. (2000). Optimizing strategic safety stock placement in supply chains. *Manufacturing and Service Operations Management*, *2*(1), 68–83 (cit. on pp. 38, 67, 69, 71).

Graves, S., & Willems, S. (2008). Strategic inventory placement in supply chains: Nonstationary demand. *Manufacturing and Service Operations Management*, *10*(2), 278–287. https://doi.org/10.1287/msom.1070.0175 (cit. on pp. 69, 70).

Graves, S., & Willems, S. (2003). Supply Chain Design: Safety Stock Placement and Supply Chain Configuration. In *Supply chain management: Design, coordination and operation* (pp. 95–132, Vol. 11). Elsevier. https://doi.org/https://doi.org/10.1016/S0927-0507(03)11003-1 (cit. on pp. 68, 69).

Graves, S., & Willems, S. (2005). Optimizing the supply chain configuration for new products. *Management Science*, *51*(8), 1165–1180. https://doi.org/10.1287/mnsc.1050.0367 (cit. on p. 72).

Greasley, A. (2009). *Operations management* (2nd). Wiley. (Cit. on p. 27).

Greenwood, P., & Nikulin, M. (1996). *A guide to chi-squared testing* (Vol. 280). John Wiley & Sons. (Cit. on p. 153).

Gröger, C., Schwarz, H., & Mitschang, B. (2014). The deep data warehouse: Link-based integration and enrichment of warehouse data and unstructured content. *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, 210–217. https://doi.org/10.1109/EDOC.2014.36 (cit. on p. 85).

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *American Statistician*, *63*, 308–319. https://doi.org/10.1198/tast.2009.08199 (cit. on pp. 151, 152).

Grubbström, R. (1998). A net present value approach to safety stocks in planned production. *International Journal of Production Economics*, *56-57*, 213–229. https://doi.org/10.1016/S0925-5273(97)00094-7 (cit. on pp. 55, 58).

Grubbström, R. (1999). Net present value approach to safety stocks in a multi-level MRP system. *International Journal of Production Economics*, *59*(1), 361–375. https://doi.org/10.1016/S0925-5273(98)00016-4 (cit. on pp. 55, 58).

Grubbström, R., & Molinder, A. (1996). Safety production plans in MRP-systems1 using transform methodology. *International Journal of Production Economics*, *46-47*, 297–309. https://doi.org/10.1016/0925-5273(95)00158-1 (cit. on pp. 54, 58).

Grubbström, R., Tang, O., Grubbstrom, R., & Tang, O. (1999). Further developments on safety stocks in an MRP system applying Laplace transforms and input-output analysis. *International Journal of Production Economics*, *60*, 381–387. https://doi.org/10.1016/S0925-5273(98)00141-8 (cit. on pp. 55, 58).

Gudehus, T. (2012). *Dynamische Disposition* (3rd ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22983-1 (cit. on p. 44).

Guide, J., & Srivastava, R. (1997). Buffering from material recovery uncertainty in a recoverable manufacturing environment. *Journal of the Operational Research Society*, *48*(5), 519–529. https://doi.org/10.1057/palgrave.jors.2600402 (cit. on pp. 60, 61).

Gunasekaran, A., Patel, C., & Tirtiroglu, E. (2001). Performance measures and metrics in a supply chain environment. *International Journal of Operations & Production Management*, *21*, 144–3577. https://www.emerald.com/insight/content/doi/10.1108/01443570110358468/full/html (cit. on p. 140).

Guyon, I., & Elisseeff, A. (2003). Special issue on variable and feature selection [cited By 67]. *Journal of Machine Learning Research*, *3* (cit. on p. 115).

Gyulai, D., Pfeiffer, A., Bergmann, J., & Gallina, V. (2018). Online lead time prediction supporting situation-aware production control. *Procedia CIRP*, *78*, 190–195. https://doi.org/10.1016/j.procir.2018.09.071 (cit. on p. 141).

Gyulai, D., Pfeiffer, A., Nick, G., Gallina, V., Sihn, W., & Monostori, L. (2018). Lead time prediction in a flow-shop environment with analytical and machine learning approaches. *IFAC PapersOnline*, *51*, 1029–1034. https://doi.org/10.1016/j.ifacol.2018.08.472 (cit. on pp. 140, 141).

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (E. Inc., Ed.; 3rd ed.). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/C2009-0-61819-5 (cit. on pp. 20, 21).

Harland, C., Brenchley, R., & Walker, H. (2003). Risk in supply networks. *Journal of Purchasing and Supply Management*, *9*(2), 51–62. https://doi.org/10.1016/S1478-4092(03)00004-9 (cit. on pp. 41, 103).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7 (cit. on pp. 117, 118, 147, 152, 153).

Hathikal, S., Chung, S., & Karczewski, M. (2020). Prediction of ocean import shipment lead time using machine learning methods. *SN Applied Sciences*, *2*. https://doi.org/10.1007/s42452-020-2951-5 (cit. on p. 141).

Hayya, J., Harrison, T., & Chatfield, D. (2009). A solution for the intractable inventory model when both demand and lead time are stochastic. *International Journal of Production Economics*, *122*(2), 595–605. https://doi.org/10.1016/j.ijpe.2009.06.001 (cit. on pp. 43, 66, 67).

Hayya, J., Harrison, T., & He, X. (2011). The impact of stochastic lead time reduction on inventory cost under order crossover. *European Journal of Operational Research*, *211*, 274–281. https://doi.org/10.1016/j.ejor.2010.11.025 (cit. on p. 137).

He, B., Huang, H., & Yuan, K. (2015). The comparison of two procurement strategies in the presence of supply disruption. *Computers and Industrial Engineering*, *85*, 296–305. https://doi.org/10.1016/j.cie.2015.03.019 (cit. on p. 43).

He, X., Xu, X., & Hayya, J. (2011). The effect of lead-time on the supply chain: The mean versus the variance. *International Journal of Information Technology and Decision Making*, *10*, 175–185. https://doi.org/10.1142/S0219622011004270 (cit. on p. 137).

Helber, S., Sahling, F., & Schimmelpfeng, K. (2013). Dynamic capacitated lot sizing with random demand and dynamic safety stocks. *OR Spectrum*, *35*(1), 75–105. https://doi.org/10.1007/s00291-012-0283-6 (cit. on pp. 46, 57, 58).

Herrmann, F. (2011). *Operative Planung in IT-Systemen für die Produktionsplanung und -steuerung* (1st ed.). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-8172-4 (cit. on p. 44).

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://misq.org/design-science-in-information-systems-research.html (cit. on p. 4).

Heydari, J., Baradaran Kazemzadeh, R., & Chaharsooghi, S. K. (2009). A study of lead time variation impact on supply chain performance. *International Journal of Advanced Manufacturing Technology*, *40*(11-12), 1206–1215. https://doi.org/10.1007/s00170-008-1428-2 (cit. on pp. 2, 137, 140).

Ho, C., Chi, Y., & Tai, Y. (2018). A Structural Approach to Measuring Uncertainty in Supply Chains. *International Journal of Electronic Commerce*, *9*(3), 91–114. https://doi.org/10.1080/10864415.2005.11044334 (cit. on p. 42).

Ho, W., Zheng, T., Yildiz, H., & Talluri, S. (2015). Supply chain risk management: A literature review. *International Journal of Production Research*, *53*(16), 5031–5069. https://doi.org/10.1080/00207543.2015.1030467 (cit. on p. 41).

Hollander, M., Wolfe, D., & Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons. (Cit. on pp. 158, 159).

Hong, Z., Dai, W., Luh, H., & Yang, C. (2018). Optimal configuration of a green product supply chain with guaranteed service time and emission constraints. *European Journal of Operational Research*, *266*(2), 663–677. https://doi.org/10.1016/j.ejor.2017.09.046 (cit. on p. 71).

Hong, Z., Lee, C., & Zhang, L. (2018). Procurement risk management under uncertainty: a review. *Industrial Management & Data Systems*, *118*(7), 1547–1574. https://doi.org/https://doi.org/10.1108/IMDS-10-2017-0469 (cit. on pp. 42, 43, 137).

Hoque, M., & Goyal, S. (2006). A heuristic solution procedure for an integrated inventory system under controllable lead-time with equal or unequal sized batch shipments between a vendor and a buyer. *International Journal of Production Economics*, *102*(2), 217–225. https://doi.org/10.1016/j.ijpe.2005.02.012 (cit. on pp. 56, 58).

Hosmer, J., D.W., Lemeshow, S., & Sturdivant, R. (2013). *Applied logistic regression: Third edition* [cited By 3437]. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118548387 (cit. on p. 117).

Hsu, H., & Wang, W. (2001). Possibilistic programming in production planning of assemble-to-order environments. *Fuzzy Sets and Systems*, *119*(1), 59–70. https://doi.org/10.1016/S0165-0114(99)00086-X (cit. on pp. 55, 58).

Hsueh, C. (2011). An inventory control model with consideration of remanufacturing and product life cycle. *International Journal of Production Economics*, *133*(2), 645–652. https://doi.org/10.1016/j.ijpe.2011.05.007 (cit. on pp. 63, 76, 78).

Hua, N. G., & Willems, S. (2016). Optimally configuring a two-stage serial line supply chain under the guaranteed service model. *International Journal of Production Economics*, *181*, 98–106. https://doi.org/10.1016/j.ijpe.2016.06.007 (cit. on pp. 69, 72, 73).

Huang, L., Song, J., & Tong, J. (2016). Supply chain planning for random demand surges: Reactive capacity and safety stock. *Manufacturing and Service Operations Management*, *18*(4), 509–524. https://doi.org/10.1287/msom.2016.0583 (cit. on pp. 56, 58, 76).

Humair, S., Ruark, J., Tomlin, B., & Willems, S. (2013). Incorporating stochastic lead times into the guaranteed service model of safety stock optimization. *Interfaces*, *43*(5), 421–434. https://doi.org/10.1287/inte.2013.0699 (cit. on pp. 72, 73).

Humair, S., & Willems, S. (2011). Optimizing strategic safety stock placement in general acyclic networks. *Operations Research*, *59*(3), 781–787. https://doi.org/10.1287/opre.1100.0913 (cit. on p. 72).

Hung, Y., & Chang, C. (1999). Determining safety stocks for production planning in uncertain manufacturing. *International Journal of Production Economics*, *58*(2), 199–208. https://doi.org/10.1016/S0925-5273(98)00124-8 (cit. on pp. 60, 61).

Iida, T. (2015). Benefits of leadtime information and of its combination with demand forecast information. *International Journal of Production Economics*, *163*, 146–156. https://doi.org/10.1016/j.ijpe.2015.02.010 (cit. on pp. 66, 67).

Inderfurth, K. (1995). Multi-stage safety stock planning with item demands correlated across products and through time. *Production and Operations Management*, *4*(2), 127–144. https://doi.org/10.1111/j.1937-5956.1995.tb00046.x (cit. on pp. 68, 70).

Inderfurth, K., & Minner, S. (1998). Safety stocks in multi-stage inventory systems under different service measures. *European Journal of Operational Research*, *106*(1), 57–73. https://doi.org/10.1016/S0377-2217(98)00210-0 (cit. on pp. 69–71).

Inderfurth, K. (2009). How to protect against demand and yield risks in MRP systems. *International Journal of Production Economics*, *121*(2), 474–481. https://doi.org/10.1016/j.ijpe.2007.02.005 (cit. on pp. 61, 62).

Inderfurth, K., & Vogelgesang, S. (2013). Concepts for safety stock determination under stochastic demand and different types of random production yield. *European Journal of Operational Research*, *224*(2), 293–301. https://doi.org/10.1016/j.ejor.2012.07.040 (cit. on p. 61).

Ivanov, T., & Pergolesi, M. (2020). The impact of columnar file formats on sql-on-hadoop engine performance: A study on orc and parquet. *Concurrency and Computation: Practice and Experience*, *32*(5), e5523 (cit. on pp. 88, 109).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer. (Cit. on p. 147).

Janssens, G., & Ramaekers, K. (2011). A linear programming formulation for an inventory management decision problem with a service constraint. *Expert Systems with Applications*, *38*(7), 7929–7934. https://doi.org/10.1016/j.eswa.2010.12.009 (cit. on pp. 57, 58, 74).

Jeong, J., Woo, J., & Park, J. (2020). Machine learning methodology for management of shipbuilding master data. *International Journal of Naval Architecture and Ocean Engineering*, *12*, 428–439. https://doi.org/10.1016/j.ijnaoe.2020.03.005 (cit. on p. 141).

Jodlbauer, H., & Reitner, S. (2012). Optimizing service-level and relevant cost for a stochastic multi-item cyclic production system. *International Journal of Production Economics*, *136*(2), 306–317. https://doi.org/10.1016/j.ijpe.2011.12.015 (cit. on pp. 55, 58).

Jonsson, P., & Mattsson, S. (2019). An inherent differentiation and system level assessment approach to inventory management: A safety stock method comparison. *International Journal of Logistics Management*, *30*(2), 663–680. https://doi.org/10.1108/IJLM-12-2017-0329 (cit. on pp. 45, 61, 75).

Juez, F. d. C., García Nieto, P., Martínez Torres, J., & Taboada Castro, J. (2010). Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model. *Mathematical and Computer Modelling*, *52*(7-8), 1177–1184. https://doi.org/10.1016/j.mcm.2010.03.017 (cit. on pp. 137, 140, 141).

Jung, J., Blau, G., Pekny, J., Reklaitis, G., & Eversdyk, D. (2008). Integrated safety stock management for multi-stage supply chains under production capacity constraints. *Computers and Chemical Engineering*, *32*(11), 2570–2581. https://doi.org/10.1016/j.compchemeng.2008.04.003 (cit. on pp. 38, 73, 137).

Jüttner, U., Peck, H., & Christopher, M. (2003). Supply chain risk management: Outlining an agenda for future research. *International Journal of Logistics: Research and Applications*, *6*(4), 197–210 (cit. on p. 41).

Kaklauskas, A. (2015). Chapter 2: Intelligent Decision Support Systems. In *Biometric and intelligent decision making support* (1st ed., p. 220, Vol. 81). Springer International Publishing. https://doi.org/10.1007/978-3-319-13659-2 (cit. on p. 19).

Kaminsky, P., & Kaya, O. (2008). Inventory positioning, scheduling and lead-time quotation in supply chains. *International Journal of Production Economics*, *114*(1), 276–293. https://doi.org/10.1016/j.ijpe.2008.02.006 (cit. on pp. 70, 71).

Kanet, J., Gorman, M., & Stosslein, M. (2010). Dynamic planned safety stocks in supply networks. *International Journal of Production Research*, *48*(22), 6859–6880. https://doi.org/10.1080/00207540903341887 (cit. on pp. 2, 60, 61, 137, 138).

Kanyalkar, A., & Adil, G. (2009). Determining the optimum safety stock under rolling schedules for capacitated multi-item production systems. *International Journal of Services and Operations Management*, *5*(4), 498–519. https://doi.org/10.1504/IJSOM.2009.024582 (cit. on pp. 55, 58).

Katircioglu, K., Brown, T., & Asghar, M. (2007). An SQL-based cost-effective inventory optimization solution. *IBM Journal of Research and Development*, *51*(3-4), 433–445. https://doi.org/10.1147/rd.513.0433 (cit. on p. 61).

Keskin, G., Omurca, S., Aydin, N., & Ekinci, E. (2015). A comparative study of production-inventory model for determining effective production quantity and safety stock level. *Applied Mathematical Modelling*, *39*(20), 6359–6374. https://doi.org/10.1016/j.apm.2015.01.037 (cit. on p. 61).

Khan, N., Ali, Z., Ali, A., McClean, S., Charles, D., Taylor, P., & Nauck, D. (2019). A generic model for end state prediction of business processes towards target compliance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11927 LNAI*, 325–335. https://doi.org/10.1007/978-3-030-34885-4_25 (cit. on pp. 106, 108).

Kim, C., Jun, J., Baek, J., Smith, R., & Kim, Y. (2005). Adaptive inventory control models for supply chain management. *International Journal of Advanced Manufacturing Technology*, *26*(9-10), 1184–1192. https://doi.org/10.1007/s00170-004-2069-8 (cit. on pp. 63, 64).

Kim, J., & Benton, W. (1995). Lot size dependent lead times in a Q, R inventory system. *International Journal of Production Research*, *33*(1), 41–58. https://doi.org/10.1080/00207549508930136 (cit. on p. 63).

Kim, J., Shin, K., & Ahn, S. (2003). A multiple replenishment contract with ARIMA demand processes. *Journal of the Operational Research Society*, *54*(11), 1189–1197. https://doi.org/10.1057/palgrave.jors.2601620 (cit. on p. 63).

Kiran, M., Murphy, P., Monga, I., Dugan, J., & Baveja, S. S. (2015). Lambda architecture for cost-effective batch and speed big data processing. *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2785–2792. https://doi.org/10.1109/BigData.2015.7364082 (cit. on p. 85).

Klawonn, F., Höppner, F., & May, S. (2011). An alternative to roc and auc analysis of classifiers. In J. Gama, E. Bradley, & J. Hollmén (Eds.), *Advances in intelligent data analysis x* (pp. 210–221). Springer Berlin Heidelberg. (Cit. on p. 104).

Klosterhalfen, S., Kallrath, J., & Fischer, G. (2014). Rail car fleet design: Optimization of structure and size. *International Journal of Production Economics*, *157*(1), 112–119. https://doi.org/10.1016/j.ijpe.2013.05.008 (cit. on pp. 55, 58).

Klosterhalfen, S., Minner, S., & Willems, S. (2014). Strategic safety stock placement in supply networks with static dual supply. *Manufacturing and Service Operations Management*, *16*(2), 204–219. https://doi.org/10.1287/msom.2013.0472 (cit. on pp. 69, 70).

Knight, F. (1921). *Risk, Uncertainty, and Profit*. Houghton Mifflin. (Cit. on p. 40).

Koch, R., CMA, & PMP. (2015). From Business Intelligence to Predictive Analytics. *Stractegic Finance Magazine*, 56–57. https://sfmagazine.com/wp-content/uploads/sfarchive/2015/01/TECH-PRACTICES-From-Business-Intelligence-to-Predictive-Analytics.pdf (cit. on pp. xiii, 19–21).

Kostrzewski, M., Varjan, P., & Gnap, J. (2020). Solutions dedicated to internal logistics 4.0. In *Sustainable logistics and production in industry 4.0* (pp. 243–262). Springer. (Cit. on p. 83).

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3 (cit. on p. 116).

Kourentzes, N., Trapero, J., & Barrow, D. (2020). Optimising forecasting models for inventory planning. *International Journal of Production Economics*, *225*, 107597 (cit. on p. 152).

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689–702. https://doi.org/10.1016/j.ejor.2016.10.031 (cit. on p. 117).

Kristianto, Y., & Zhu, L. (2013). An integration of assembly planning by design into supply chain planning. *International Journal of Advanced Manufacturing Technology*, *69*(5-8), 1593–1604. https://doi.org/10.1007/s00170-013-5060-4 (cit. on p. 73).

Kristianto, Y., Gunasekaran, A., Helo, P., & Sandhu, M. (2012). A decision support system for integrating manufacturing and product design into the reconfiguration of the supply chain networks. *Decision Support Systems*, *52*(4), 790–801. https://doi.org/10.1016/j.dss.2011.11.014 (cit. on p. 73).

Krupp, J. (1997). Safety stock management. *Production and Inventory Management Journal*, *38*(3), 11–18 (cit. on pp. 56, 58).

Kumar, A., & Evers, P. (2015). Setting safety stock based on imprecise records. *International Journal of Production Economics*, *169*, 68–75. https://doi.org/10.1016/j.ijpe.2015.07.018 (cit. on p. 61).

Kumar, D., & Kumar, D. (2018). Managing the essential medicines stock at rural healthcare systems in India. *International Journal of Health Care Quality Assurance*, *31*(8), 950–965. https://doi.org/10.1108/IJHCQA-12-2016-0186 (cit. on p. 73).

Kumar, K., & Aouam, T. (2018a). Effect of setup time reduction on supply chain safety stocks. *Journal of Manufacturing Systems*, *49*, 1–15. https://doi.org/10.1016/j.jmsy.2018.08.001 (cit. on p. 71).

Kumar, K., & Aouam, T. (2018b). Integrated lot sizing and safety stock placement in a network of production facilities. *International Journal of Production Economics*, *195*, 74–95. https://doi.org/10.1016/j.ijpe.2017.10.006 (cit. on pp. 38, 67, 71).

Kumar, K., & Aouam, T. (2019). Extending the strategic safety stock placement model to consider tactical production smoothing. *European Journal of Operational Research*, *279*(2), 429–448. https://doi.org/10.1016/j.ejor.2019.06.009 (cit. on p. 69).

Kumar, M., Garg, D., & Agarwal, A. (2019). Cause and effect analysis of inventory management in leagile supply chain. *Journal of Management Information and Decision Science*, *22*(2), 67–100 (cit. on p. 71).

Lambert, D., Stock, J., & Ellram, L. (1998). *Fundamentals of Logistics Management* (McGraw-Hill Companies Inc., Ed.; 1st ed.). Gary Burke. (Cit. on pp. 26, 28, 29, 39, 42, 43).

Larose, D. (2005). *Discovering knowledge in data: An introduction to data mining*. Wiley Blackwell. https://doi.org/10.1002/0471687545 (cit. on pp. 116, 118, 121, 151).

Lee, C., & Rim, S. (2019). A Mathematical Safety Stock Model for DDMRP Inventory Replenishment. *Mathematical Problems in Engineering*, *2019*. https://doi.org/10.1155/2019/6496309 (cit. on pp. 57, 58, 74).

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access*, *5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365 (cit. on pp. 88, 109).

Li, C., Erlebacher, S., & Kropp, D. (1997). Investment in setup cost, lead time, and demand predictability improvement in the EOQ model. *Production and Operations Management*, *6*(4), 341–352 (cit. on p. 58).

Li, H., & Jiang, D. (2012). New model and heuristics for safety stock placement in general acyclic supply chain networks. *Computers and Operations Research*, *39*(7), 1333–1344. https://doi.org/10.1016/j.cor.2011.08.001 (cit. on pp. xiv, 38, 67, 68).

Li, S., Li, X., Zhang, D., & Zhou, L. (2017). Joint optimization of distribution network design and two-echelon inventory control with stochastic demand and CO2 emission tax charges. *PLoS One*, *12*(1). https://doi.org/10.1371/journal.pone.0168526 (cit. on pp. 70, 71).

Li, S., & Zhang, X. (2020). Research on orthopedic auxiliary classification and prediction model based on xgboost algorithm. *Neural Computing and Applications*, *32*, 1971–1979 (cit. on p. 116).

Li, Z., Fei, W., Zhou, E., Gajpal, Y., & Chen, X. (2019). The impact of lead time uncertainty on supply chain performance considering carbon cost. *Sustainability (Switzerland)*, *11*(22), 1–19. https://doi.org/10.3390/su11226457 (cit. on pp. 2, 137).

Lian, Z., Deshmukh, A., & Wang, J. (2006). The optimal frozen period in a dynamic production model. *International Journal of Production Economics*, *103*(2), 648–655. https://doi.org/10.1016/j.ijpe.2005.12.005 (cit. on pp. 63, 64).

Liao, S., Hsieh, C., & Lin, Y. (2011). A multi-objective evolutionary optimization approach for an integrated location-inventory distribution network problem under vendor-managed inventory systems. *Annals of Operations Research*, *186*(1), 213–229. https://doi.org/10.1007/s10479-010-0801-3 (cit. on pp. 70, 71).

Liao, Y., Deschamps, F., Loures, E. d. F. R., & Ramos, L. F. P. (2017). Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal. *International Journal of Production Research*, *55*(12), 3609–3629. https://doi.org/10.1080/00207543.2017.1308576 (cit. on p. 15).

Lim, H., Umi, Y., & Haziqah, S. (2019). Manufacturing lead time classification using support vector machine (H. B. Zaman, A. F. Smeaton, T. K. Shih, S. Velastin, T. Terutoshi, N. M. Ali, & M. N. Ahmad, Eds.). *International Visual Informatics Conference*, *11870*, 268–278. https://doi.org/10.1007/978-3-030-34032-2 (cit. on p. 141).

Lin, G., Breitwieser, R., Cheng, F., Eagen, J., & Ettl, M. (2000). Product hardware complexity and its impact on inventory and customer on-time delivery. *International Journal of Flexible Manufacturing Systems*, *12*(2-3), 145–163. https://doi.org/10.1023/A:1008191530004 (cit. on p. 73).

Lingitz, L., Gallina, V., Ansari, F., Gyulai, D., Pfeiffer, A., & Sihn, W. (2018). Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. *Procedia CIRP*, *72*, 1051–1056. https://doi.org/10.1016/j.procir.2018.03.148 (cit. on p. 141).

Liu, D., & Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, *45*, 503–528 (cit. on p. 158).

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/https://doi.org/10.1002/widm.8 (cit. on p. 151).

Louly, M., & Dolgui, A. (2009). Calculating safety stocks for assembly systems with random component procurement lead times: A branch and bound algorithm. *European Journal of Operational Research*, *199*(3), 723–731. https://doi.org/https://doi.org/10.1016/j.ejor.2007.11.066 (cit. on pp. 2, 57, 59, 137, 138).

Louly, M. A., Dolgui, A., & Hnaien, F. (2008). Supply planning for single-level assembly system with stochastic component delivery times and service-level constraint. *International Journal of Production Economics*, *115*(1), 236–247. https://doi.org/10.1016/j.ijpe.2008.06.005 (cit. on pp. 65, 75, 78).

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on pp. iii, iv).

Lu, H., Wang, H., Xie, Y., & Li, H. (2016). Construction material safety-stock determination under nonstationary stochastic demand and random supply yield. *IEEE Transactions on Engineering Management*, *63*(2), 201–212. https://doi.org/10.1109/TEM.2016.2536146 (cit. on pp. 61, 62).

Lueth, K., Patsioura, C., Williams, Z., & Kermani, Z. (2016). *Industrial Analytics 2016/2017: The current state of data analytics usage in industrial companies* (tech. rep.). Digital Analytics Association e.V. Germany (DAAG). https://iot-analytics.com/wp/wp-content/uploads/2016/10/Industrial-Analytics-Report-2016-2017-vp-singlepage.pdf (cit. on pp. 23, 103, 168).

Lukinskiy, V., & Lukinskiy, V. (2017). Evaluation of stock management strategies reliability at dependent demand. *Transport and Telecommunication*, *18*(1), 60–69. https://doi.org/10.1515/ttj-2017-0006 (cit. on pp. 57, 58).

Luo, C. (2022). A comparison analysis for credit scoring using bagging ensembles. *Expert Systems*, *39*. https://doi.org/10.1111/exsy.12297 (cit. on pp. 116, 118).

Manary, M., & Willems, S. (2008). Setting safety-stock targets at Intel in the presence of forecast bias. *Interfaces*, *38*(2), 112–122. https://doi.org/10.1287/inte.1070.0339 (cit. on pp. 69, 70).

Manary, M., Willems, S., & Shihata, A. (2009). Correcting heterogeneous and biased forecast error at intel for supply chain optimization. *Interfaces*, *39*(5), 415–427. https://doi.org/10.1287/inte.1090.0452 (cit. on pp. 69, 70).

Manavalan, E., & Jayakrishna, K. (2019). A review of internet of things (iot) embedded sustainable supply chain for industry 4.0 requirements [cited By 160]. *Computers and Industrial Engineering*, *127*, 925–953. https://doi.org/10.1016/j.cie.2018.11.030 (cit. on p. 103).

Manuj, I., & Mentzer, J. (2008). Global supply chain risk management strategies. *International Journal of Physical Distribution and Logistics Management*, *38*(3), 192–223. https://doi.org/10.1108/09600030810866986 (cit. on p. 41).

March, J., & Shapira, Z. (1987). Managerial Perspectives on Risk and Risk Taking. *Management Science*, *33*(1), 1404–1418 (cit. on p. 41).

Martinelli, F., & Valigi, P. (2004). Hedging point policies remain optimal under limited backlog and inventory space. *IEEE Transactions on Automatic Control*, *49*(10), 1863–1869. https://doi.org/10.1109/TAC.2004.835592 (cit. on p. 60).

McNair, D. (2015). Enhancing Nursing Staffing Forecasting with Safety Stock over Lead Time Modeling. *Nursing Administration Quarterly*, *39*(4), 291–296. https://doi.org/10.1097/NAQ.0000000000000124 (cit. on pp. 60, 61).

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R., Zaharia, M., & Talwalkar, A. (2016). Mllib: Machine learning in apache spark. *J. Mach. Learn. Res.*, *17*(1), 1235–1241 (cit. on pp. 89, 109).

Michna, Z., Nielsen, P., & Nielsen, I. (2018). The impact of stochastic lead times on the bullwhip effect–a theoretical insight. *Production and Manufacturing Research*, *6*, 190–200. https://doi.org/10.1080/21693277.2018.1484822 (cit. on p. 137).

Minner, S. (1997). Dynamic programming algorithms for multi-stage safety stock optimization. *OR Spectrum*, *19*(4), 261–271. https://doi.org/10.1007/BF01539783 (cit. on pp. 69, 70).

Mitchell, V. (1995). Organizational Risk Perception and Reduction: A Literature Review. *British Journal of Management*, *6*(2), 115–133. https://doi.org/10.1111/j.1467-8551.1995.tb00089.x (cit. on p. 41).

Moeeni, F., Replogle, S., Chaudhury, Z., & Syamil, A. (2012). A refinement of the classical order point model. *International Journal of Information Systems and Supply Chain Management*, *5*(3), 43–57. https://doi.org/10.4018/jisscm.2012070103 (cit. on pp. 56, 58).

Molinder, A. (1997). Joint optimization of lot-sizes, safety stocks and safety lead times in an MRP system. *International Journal of Production Research*, *35*(4), 983–994. https://doi.org/10.1080/002075497195498 (cit. on pp. 60, 61, 78).

Moncayo-Martinez, L., Resendiz-Flores, E., Mercado, D., & Sanchez-Ramirez, C. (2014). Placing Safety Stock in Logistic Networks under Guaranteed-Service Time Inventory Models: An Application to the Automotive Industry. *Journal of Applied Research and Technology*, *12*(3), 538–550. https://doi.org/10.1016/S1665-6423(14)71633-5 (cit. on pp. 69, 70).

Moncayo–Martínez, L., Ramírez–López, A., & Recio, G. (2016). Managing inventory levels and time to market in assembly supply chains by swarm intelligence algorithms. *International Journal of Advanced Manufacturing Technology*, *82*(1-4), 419–433. https://doi.org/10.1007/s00170-015-7313-x (cit. on pp. 69, 70).

Moncayo-Martínez, L., & Zhang, D. (2013). Optimising safety stock placement and lead time in an assembly supply chain using bi-objective MAX-MIN ant system. *International Journal of Production Economics*, *145*(1), 18–28. https://doi.org/10.1016/j.ijpe.2012.12.024 (cit. on pp. 69–71).

Monczka, R., Handfield, R., Giunipero, L., Patterson, J., & Waters, D. (2010). *Purchasing & Supply Chain Management*. Cengage Learning EMEA. (Cit. on p. 39).

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5 (cit. on p. 46).

Monostori, L., Kumara, S., Sauer, O., Ueda, K., Reinhart, G., Kádár, B., Schuh, G., Bauernhansl, T., Sihn, W., & Kondoh, S. (2016). Cyber-physical systems in manufacturing. *CIRP Annals*, *65*(2), 621–641. https://doi.org/10.1016/j.cirp.2016.06.005 (cit. on p. 17).

Monthatipkul, C., Das, S., & Yenradee, P. (2010). Distribution policy in an M-store regional supply chain. *International Journal of Integrated Supply Management*, *5*(3), 214–238. https://doi.org/10.1504/IJISM.2010.033976 (cit. on p. 70).

Monthatipkul, C., & Yenradee, P. (2007). Positioning safety stock in a one-warehouse multi-retailer supply chain controlled by optimal inventory/distribution plan. *International Journal of Industrial Engineering - Theory Applications and Practice*, *14*(2), 169–178 (cit. on p. 70).

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31 (cit. on p. 151).

Mou, Q., Cheng, Y., & Liao, H. (2017). A note on "lead time reduction strategies in a single-vendor-single-buyer integrated inventory model with lot size-dependent lead times and stochastic demand". *International Journal of Production Economics*, *193*, 827–831. https://doi.org/10.1016/j.ijpe.2017.09.012 (cit. on pp. 56, 58).

Nasiri, G., Davoudpour, H., & Karimi, B. (2010). The impact of integrated analysis on supply chain management: A coordinated approach for inventory control policy. *Supply Chain Management*, *15*(4), 277–289. https://doi.org/10.1108/13598541011054652 (cit. on pp. 70, 71).

NBD-PWG. (2015). *Nist big data interoperability framework: Volume 6, reference architecture* (Technical Report No. NIST SP 1500-6). National Institute of Standards and Technology. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6.pdf (cit. on p. 85).

Neale, J., & Willems, S. (2009). Managing inventory in supply chains with nonstationary demand. *Interfaces*, *39*(5), 388–399. https://doi.org/10.1287/inte.1090.0442 (cit. on pp. 69, 70).

Negahban, A., & Dehghanimohammadabadi, M. (2018). Optimizing the supply chain configuration and production-sales policies for new products over multiple planning horizons. *International Journal* *of Production Economics*, *196*, 150–162. https://doi.org/10.1016/j.ijpe.2017.11.019 (cit. on p. 71).

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Source: Journal of the Royal Statistical Society. Series A (General)*, *135*, 370 (cit. on p. 152).

Nenni, M., & Schiraldi, M. (2013). Validating virtual safety stock effectiveness through simulation. *International Journal of Engineering Business Management*, *5*(1). https://doi.org/10.5772/56859 (cit. on p. 42).

Ngo, V., Le-Khac, N.-A., & Kechadi, M.-T. (2019). Designing and implementing data warehouse for agricultural big data. In K. Chen, S. Seshadri, & L.-J. Zhang (Eds.), *Big data – bigdata 2019* (pp. 1–17). Springer International Publishing. (Cit. on pp. 84, 85).

Niemi, T., Hameri, A., Kolesnyk, P., & Appelqvist, P. (2020). What is the value of delivering on time? *Journal of Advances in Management Research*, *17*, 473–503. https://doi.org/10.1108/JAMR-12-2019-0218 (cit. on p. 104).

North, M. (2012). *Data Mining for the Masses*. (Cit. on pp. 7, 8).

Ohno, K., Nakashima, K., & Kojima, M. (1995). Optimal numbers of two kinds of kanbans in a JIT production system. *International Journal of Production Research*, *33*(5), 1387–1401. https://doi.org/10.1080/00207549508930216 (cit. on pp. 57, 58).

Olesków-Szłapka, J., & Stachowiak, A. (2019). The framework of logistics 4.0 maturity model. In A. Burduk, E. Chlebus, T. Nowakowski, & A. Tubis (Eds.), *Intelligent systems in production engineering and maintenance* (pp. 771–781). Springer International Publishing. (Cit. on p. 83).

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Syst. Appl.*, *73*, 125–144. https://doi.org/10.1016/j.eswa.2016.12.036 (cit. on pp. xiv, xv, 119, 154, 155).

Osman, H., & Demirli, K. (2012). Integrated safety stock optimization for multiple sourced stockpoints facing variable demand and lead time. *International Journal of Production Economics*, *135*(1), 299–307. https://doi.org/10.1016/j.ijpe.2011.08.004 (cit. on pp. 72, 73).

Ozguven, E., & Ozbay, K. (2012). Case study-based evaluation of stochastic multicommodity emergency inventory management model. *Transportation Research Record*, (2283), 12–24. https://doi.org/10.3141/2283-02 (cit. on pp. 66, 67).

Ozsen, L., Coullard, C., & Daskin, M. (2008). Capacitated warehouse location model with risk pooling. *Naval Research Logistics*, *55*(4), 295–312. https://doi.org/10.1002/nav.20282 (cit. on pp. 70, 71).

Öztürk, A., Kayalıgil, S., & Özdemirel, N. (2006). Manufacturing lead time estimation using data mining. *European Journal of Operational Research*, *173*(2), 683–700. https://doi.org/https://doi.org/10.1016/j.ejor.2005.03.015 (cit. on pp. 140, 141).

Panetto, H., Iung, B., Ivanov, D., Weichhart, G., & Wang, X. (2019). Challenges for the cyber-physical manufacturing enterprises of the future. *Annual Reviews in Control*, *47*, 200–213 (cit. on p. 83).

Peck, H. (2006). Reconciling supply chain vulnerability, risk and supply chain management. *International Journal of Logistics Research and Applications*, *9*(2), 127–142. https://doi.org/10.1080/13675560600673578 (cit. on p. 41).

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302 (cit. on pp. xiii, 4–6).

Persona, A., Battini, D., Manzini, R., & Pareschi, A. (2007). Optimal safety stock levels of subassemblies and manufacturing components. *International Journal of Production Economics*, *110*(1-2), 147–159. https://doi.org/10.1016/j.ijpe.2007.02.020 (cit. on pp. 55, 58).

Petridis, K. (2015). Optimal design of multi-echelon supply chain networks under normally distributed demand. *Annals of Operations Research*, *227*(1), 63–91. https://doi.org/10.1007/s10479-013-1420-6 (cit. on pp. 70, 71).

Prak, D., & Teunter, R. (2019). A general method for addressing forecasting uncertainty in inventory models. *International Journal of Forecasting*, *35*(1), 224–238. https://doi.org/10.1016/j.ijforecast.2017.11.004 (cit. on p. 174).

Prak, D., Teunter, R., & Syntetos, A. (2017). On the calculation of safety stocks when demand is forecasted. *European Journal of Operational Research*, *256*(2), 454–461. https://doi.org/10.1016/j.ejor.2016.06.035 (cit. on pp. 56, 58).

Prawira, A., Yuliani, E., & Iridiastadi, H. (2019). Proposed inventory strategy of NSR material in Cikarang-Indonesia oil and gas: Services company. *Jordan Journal of Mechanical and Industrial Engineering*, *12*(3), 179–188 (cit. on pp. 55, 58).

Qin, X., Chen, Y., Chen, J., Li, S., Liu, J., & Zhang, H. (2017). The performance of sql-on-hadoop systems - an experimental study. *2017 IEEE International Congress on Big Data (BigData Congress)*, 464–471. https://doi.org/10.1109/BigDataCongress.2017.68 (cit. on p. 89).

Quah, J., Ang, C., Divakar, R., Idrus, R., Abdullah, N., & Chew, X. (2019). Timing-of-delivery prediction model to visualize delivery trends for pos laju malaysia by machine learning techniques [cited By 0]. *Communications in Computer and Information Science*, *937*, 85–95. https://doi.org/10.1007/978-981-13-3441-2_7 (cit. on pp. 106, 108).

Rao, S., & Goldsby, T. (2009). Supply chain risks: A review and typology. *The International Journal of Logistics Management*, *20*(1), 97–123. https://doi.org/10.1108/09574090910954864 (cit. on p. 41).

Rappold, J., & Yoho, K. (2014). Setting safety stocks for stable rotation cycle schedules. *International Journal of Production Economics*, *156*, 146–158. https://doi.org/10.1016/j.ijpe.2014.05.020 (cit. on pp. 55, 58).

Reichhart, A., Framinan, J., & Holweg, M. (2008). On the link between inventory and responsiveness in multi-product supply chains. *International Journal of Systems Science*, *39*(7), 677–688. https://doi.org/10.1080/00207720802090856 (cit. on p. 58).

Ribeiro, D., Matos, L. M., Moreira, G., Pilastri, A., & Cortez, P. (2022). Isolation forests and deep autoencoders for industrial screw tightening anomaly detection. *Computers*, *11*. https://doi.org/10.3390/computers11040054 (cit. on p. 8).

Richardson, J., Schlegel, K., Sallam, R., Kronz, A., & Sun, J. (2021). 2021 gartner magic quadrant for analytics and business intelligence platforms. https://www.gartner.com/doc/reprints?id=1-254T1IQX&ct=210202&st=sb (cit. on p. 111).

Riezebos, J. (2006). Inventory order crossovers. *International Journal of Production Economics*, *104*(2), 666–675. https://doi.org/10.1016/j.ijpe.2004.11.011 (cit. on pp. 75, 174).

Ross, A., Khajehnezhad, M., Otieno, W., & Aydas, O. (2017). Integrated location-inventory modelling under forward and reverse product flows in the used merchandise retail sector: A multi-echelon formulation. *European Journal of Operational Research*, *259*(2), 664–676. https://doi.org/10.1016/j.ejor.2016.10.036 (cit. on p. 71).

Roßmann, B., Canzaniello, A., von der Gracht, H., & Hartmann, E. (2018). The future and social impact of Big Data Analytics in Supply Chain Management: Results from a Delphi study. *Technological Forecasting and Social Change*, *130*, 135–149. https://doi.org/10.1016/j.techfore.2017.10.005 (cit. on p. 103).

Royal Society. (1992). *Risk : analysis, perception and management*. (Cit. on p. 41).

Ruiz-Torres, A., & Mahmoodi, F. (2010). Safety stock determination based on parametric lead time and demand information. *International Journal of Production Research*, *48*(10), 2841–2857. https://doi.org/10.1080/00207540902795299 (cit. on pp. 2, 44, 45, 66, 67, 74, 138, 157).

Rushton, A., Croucher, P., & Baker, P. (2014). *The Handbook of Logistics and Distribution Management* (5th ed.). Kogan Page Limited. (Cit. on pp. xiii, 26–30, 39, 40, 43).

Russel, S., & Norving, P. (2010). *Artificial Intelligence: A Modern Approach* (P. E. Inc., Ed.; 3rd ed.). https://doi.org/10.1017/S0269888900007724 (cit. on pp. 21, 151).

Saad, S., Merino Perez, C., & Vega Alvarado, V. (2017). Development of a mechanism to facilitate the safety stock planning configuration in ERP. *Production and Manufacturing Research*, *5*(1), 42–56. https://doi.org/10.1080/21693277.2017.1322541 (cit. on pp. 60, 61).

Saad, S., Perez, C., & Alvarado, V. (2017). Development of a mechanism to facilitate the safety stock planning configuration in erp. *Production and Manufacturing Research*, *5*, 42–56. https://doi.org/10.1080/21693277.2017.1322541 (cit. on pp. 2, 138).

Sakulsom, N., & Tharmmaphornphilas, W. (2019). Heuristics for a periodic-review policy in a two-echelon inventory problem with seasonal demand. *Computers & Industrial Engineering*, *133*, 292–302. https://doi.org/10.1016/j.cie.2019.05.017 (cit. on pp. 63, 64).

Sana, S., & Chaudhuri, K. (2010). An EMQ model in an imperfect production process. *International Journal of Systems Science*, *41*(6), 635–646. https://doi.org/10.1080/00207720903144495 (cit. on pp. 59, 60).

Sanders, N. (2014). *The definitive guide to manufacturing and service operations* (Council of Supply Chain Management Professionals, Ed.; 1st ed.). Pearson Education. (Cit. on pp. 27, 28).

Sanders, N. R. (2012). *Supply Chain Management: A Global Perspective*. John Wiley & Sons, Inc. (Cit. on pp. 27, 39, 40, 43).

Santos, M., & Costa, C. (2020). Big data: Concepts, warehousing, and analytics. In *Big data: Concepts, warehousing, and analytics* (pp. 1–284). River Publishers. (Cit. on pp. 85, 91, 92).

Santos, M., Andrade, C., Costa, C., Martinho, B., Costa, E., Galvão, J., & Lima, F. (2017). Evaluating sql-on-hadoop for big data warehousing on not-so-good hardware. *IDEAS 2017: Proceedings of the 21st International Database Engineering & Applications Symposium, Part F129476*, 242–252 (cit. on pp. 111, 145).

Santos, M. Y., Martinho, B., & Costa, C. (2017). Modelling and implementing big data warehouses for decision support. *Journal of Management Analytics*, *4*(2), 111–129. https://doi.org/10.1080/23270012.2017.1304292 (cit. on p. 85).

Santos, M. Y., Oliveira e Sá, J., Andrade, C., Vale Lima, F., Costa, E., Costa, C., Martinho, B., & Galvão, J. (2017). A big data system supporting bosch braga industry 4.0 strategy. *International Journal of Information Management*, *37*(6), 750–760. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2017.07.012 (cit. on p. 83).

Santoso, L., & Yulia. (2017). Data warehouse with big data technology for higher education [4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia]. *Procedia Computer Science*, *124*, 93–99. https://doi.org/https://doi.org/10.1016/j.procs.2017.12.134 (cit. on p. 84).

Sarkar, M., & Sarkar, B. (2019). Optimization of Safety Stock under Controllable Production Rate and Energy Consumption in an Automated Smart Production Management. *Energies*, *12*(11). https://doi.org/10.3390/en12112059 (cit. on pp. 59, 60).

Scarf, H. (1958). Stationary operating characteristics of an inventory model with time lag. *Studies in the mathematical theory of inventory and production*, 298–319 (cit. on pp. 137, 174).

Schmidt, M., Hartmann, W., & Nyhuis, P. (2012). Simulation based comparison of safety-stock calculation methods. *CIRP Annals - Manufacturing Technology*, *61*(1), 403–406. https://doi.org/10.1016/j.cirp.2012.03.054 (cit. on pp. xvi, 38, 43–45, 75).

Schneider, H., & Rinks, P., D.and Kelle. (1995). Power approximations for a two▯echelon inventory system using service levels. *Production and Operations Management*, *4*(4), 381–400. https://doi.org/10.1111/j.1937-5956.1995.tb00300.x (cit. on pp. 70, 72).

Schoenmeyr, T., & Graves, S. (2009). Strategic safety stocks in supply chains with evolving forecasts. *Manufacturing and Service Operations Management*, *11*(4), 657–673. https://doi.org/10.1287/msom.1080.0245 (cit. on pp. 69, 70).

Schuster Puga, M., Minner, S., & Tancrez, J. (2019). Two-stage supply chain design with safety stock placement decisions. *International Journal of Production Economics*, *209*, 183–193. https://doi.org/10.1016/j.ijpe.2018.05.018 (cit. on p. 73).

Schuster Puga, M., & Tancrez, J. (2017). A heuristic algorithm for solving large location–inventory problems with demand uncertainty. *European Journal of Operational Research*, *259*(2), 413–423. https://doi.org/10.1016/j.ejor.2016.10.037 (cit. on p. 71).

Sebaa, A., Chikh, F., Nouicer, A., & Tari, A. (2018). Medical big data warehouse: Architecture and system design, a case study: Improving healthcare resources distribution. *Journal of medical systems*, *42*(4), 59. https://doi.org/10.1007/s10916-018-0894-9 (cit. on pp. 84, 85).

Sellitto, M. (2018). Lead-time, inventory, and safety stock calculation in job-shop manufacturing. *Acta Polytechnica*, *58*(6), 395–401. https://doi.org/10.14311/AP.2018.58.0395 (cit. on pp. 57, 59).

Shahabi, M., Tafreshian, A., Unnikrishnan, A., & Boyles, S. (2018). Joint production-inventory-location problem with multi-variate normal demand. *Transportation Research Part B - Methodological*, *110*, 60–78. https://doi.org/10.1016/j.trb.2018.02.002 (cit. on p. 71).

Shang, K. (2012). Single-stage approximations for optimal policies in serial inventory systems with non-stationary demand. *Manufacturing and Service Operations Management*, *14*(3), 414–422. https://doi.org/10.1287/msom.1110.0373 (cit. on pp. 57, 58).

Sharda, R., Delen, D., & Turban, E. (2015). *Business intelligence and Analytics: Systems for Decision Support* (10th ed.). New Jersey 07458. https://doi.org/10.4324/9781315206455-12 (cit. on pp. 19, 20).

Shekhar, S., Bansode, A., & Salim, A. (2021). A comparative study of hyper-parameter optimization tools. *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–6 (cit. on p. 154).

Shen, Z., Coullard, C., & Daskin, M. (2003). A joint location-inventory model. *Transportation Science*, *37*(1), 40–55. https://doi.org/10.1287/trsc.37.1.40.12823 (cit. on pp. 70, 73).

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly: : Management Information Systems*, *35*(3), 553–572. https://doi.org/10.2139/ssrn.1606674 (cit. on p. 21).

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The hadoop distributed file system. *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1–10. https://doi.org/10.1109/MSST.2010.5496972 (cit. on pp. 88, 108).

Silva, N., Barros, J., Santos, M., Costa, C., Cortez, P., Carvalho, M. S., & Gonçalves, J. (2021). Advancing logistics 4.0 with the implementation of a big data warehouse: A demonstration case for the automotive industry. *Electronics (Switzerland)*, *10* (cit. on pp. xv, 104, 111, 138, 139, 142, 144–146).

Silva, P., Gonçalves, J., TiagoMartins, Marques, L., Oliveira, M., Reis, M., Araújo, L., Correia, D., Telhada, J., Costa, L., & Fernandes, J. (2022). A hybrid bi-objective optimization approach for joint determination of safety stock and safety time buffers in multi-item single-stage industrial supply chains. *Computers & Industrial Engineering*, *168*, 108095. https://doi.org/https://doi.org/10.1016/j.cie.2022.108095 (cit. on p. 174).

Silver, E., Pyke, D., & Thomas, D. (2016). *Inventory and production management in supply chains* (4th ed.). https://doi.org/10.1201/9781315374406 (cit. on pp. 137, 142, 152, 174).

Simchi-Levi, D., & Zhao, Y. (2005). Safety stock positioning in supply chains with stochastic lead times. *Manufacturing and Service Operations Management*, *7*(4), 295–318. https://doi.org/10.1287/msom.1050.0087 (cit. on pp. 72, 73).

Simchi-Levi, D., Kamimsky, P., & Simchi-Levi, E. (2000). *Designing and managing the Supply Chain: Concepts, Strategies, and Case Studies* (McGraw-Hill Companies, Ed.; 1st ed.). (Cit. on pp. 28, 30, 103, 136).

Simpson, K. (1958). In-Process Inventories. *Operations Research*, *6*(6), 791–908. https://doi.org/https://doi.org/10.1287/opre.6.6.863 (cit. on p. 68).

Singh, S., & Soni, U. (2019). Predicting order lead time for just in time production system using various machine learning algorithms: A case study. *International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 1–4 (cit. on pp. 137, 140–142).

Sitompul, C., Aghezzaf, E., Dullaert, W., & Van Landeghem, H. (2008). Safety stock placement problem in capacitated supply chains. *International Journal of Production Research*, *46*(17), 4709–4727. https://doi.org/10.1080/00207540701278406 (cit. on pp. 68–70).

Sodhi, M., Son, B., & Tang, C. (2012). Researchers' perspectives on supply chain risk management. *Production and Operations Management*, *21*, 1–13 (cit. on p. 103).

Song, Z. (2017). Determination of inventory for mining production with a real options approach and comparison with other classic methods. *International Journal of Mining, Reclamation and Environment*, *31*(5), 346–363. https://doi.org/10.1080/17480930.2016.1156871 (cit. on pp. 59, 60).

Sonntag, D., & Kiesmüller, G. (2017). The Influence of Quality Inspections on the Optimal Safety Stock Level. *Production and Operations Management*, *26*(7), 1284–1298. https://doi.org/10.1111/poms.12691 (cit. on pp. 72, 73).

Spanaki, K., Gürgüç, Z., Adams, R., & Mulligan, C. (2018). Data supply chain (dsc): Research synthesis and future directions. *International Journal of Production Research*, *56*, 4447–4466. https://doi.org/10.1080/00207543.2017.1399222 (cit. on p. 103).

Spark, A. (2018). Apache spark. *Retrieved January*, *17*, 2018 (cit. on pp. 88, 109, 111).

Spark, A. (2020). Classification and regression - Spark 3.0.1 Documentation. (Cit. on pp. xvii, 116–118, 120, 125, 151).

Srinivasan, M., Novack, R., & Thomas, D. (2011). Optimal and approximate policies for inventory systems with order crossover. *Journal of Business Logistics*, *32*(2), 180–193. https://doi.org/https://doi.org/10.1111/j.2158-1592.2011.01015.x (cit. on p. 174).

Srivastav, A., & Agrawal, S. (2016). Multi-objective optimization of hybrid backorder inventory model. *Expert Systems with Applications*, *51*, 76–84. https://doi.org/10.1016/j.eswa.2015.12.032 (cit. on pp. 56, 58).

Srivastav, A., & Agrawal, S. (2018). On a single item single stage mixture inventory models with independent stochastic lead times. *Operational Research*, 1–39. https://doi.org/10.1007/s12351-018-0408-z (cit. on p. 43).

Strandhagen, J., Vallandingham, L., Fragapane, G., Strandhagen, J. W., Stangeland, A., & Sharma, N. (2017). Logistics 4.0 and emerging sustainable business models. *Advances in Manufacturing*, *5*(4), 359–369 (cit. on pp. 83, 86).

Strohhecker, J., & Größler, A. (2019). Threshold behavior of optimal safety stock coverage in the presence of extended production disruptions. *Journal of Modelling in Management*. https://doi.org/10.1108/JM2-03-2019-0074 (cit. on pp. 60, 61, 76).

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719. https://doi.org/10.1142/S0218001409007326 (cit. on p. 121).

Sutton, C. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*, *24*, 303–329 (cit. on p. 151).

Swaminathan, J., & Tayur, S. (1998). Managing broader product lines through delayed differentiation using vanilla boxes. *Management Science*, *44*(12, 2), S161–S172. https://doi.org/10.1287/mnsc.44.12.S161 (cit. on p. 73).

Sydow, J., & Frenkel, S. (2013). Labor, risk, and uncertainty in global supply networks - Exploratory insights. *Journal of Business Logistics*, *34*(3), 236–247. https://doi.org/10.1111/jbl.12022 (cit. on p. 40).

Taleizadeh, A., Niaki, S., & Barzinpour, F. (2011). Multiple-buyer multiple-vendor multi-product multi-constraint supply chain problem with stochastic demand and variable lead-time: A harmony search algorithm. *Applied Mathematics and Computation*, *217*(22), 9234–9253. https://doi.org/10.1016/j.amc.2011.04.001 (cit. on pp. 65, 66).

Taleizadeh, A., Samimi, H., Sarkar, B., & Mohammadi, B. (2017). Stochastic machine breakdown and discrete delivery in an imperfect inventory-production system. *Journal of Industrial and Management Optimization*, *13*(3), 1511–1535. https://doi.org/10.3934/jimo.2017005 (cit. on pp. 59, 60).

Talluri, S., Cetin, K., & Gardner, A. (2004). Integrating demand and supply variability into safety stock evaluations. *International Journal of Physical Distribution and Logistics Management*, *34*(1), 62–69. https://doi.org/10.1108/09600030410515682 (cit. on pp. 2, 60, 61, 138).

Tang, C. (2006a). Perspectives in supply chain risk management. *International Journal of Production Economics*, *103*(2), 451–488. https://doi.org/10.1016/j.ijpe.2005.12.006 (cit. on p. 41).

Tang, C. (2006b). Perspectives in supply chain risk management. *International Journal of Production Economics*, *103*, 451–488. https://doi.org/10.1016/j.ijpe.2005.12.006 (cit. on p. 103).

Tang, C., & Veelenturf, L. (2019). The strategic role of logistics in the industry 4.0 era. *Transportation Research Part E: Logistics and Transportation Review*, *129*, 1–11. https://doi.org/https://doi.org/10.1016/j.tre.2019.06.004 (cit. on p. 83).

Tang, L., Liu, G., & Liu, J. (2008). Raw material inventory solution in iron and steel industry using Lagrangian relaxation. *Journal of the Operational Research Society*, *59*(1), 44–53. https://doi.org/10.1057/palgrave.jors.2602335 (cit. on pp. 65, 66).

Tang, O., & Musa, S. (2011). Identifying risk issues and research advancements in supply chain risk management. *International Journal of Production Economics*, *133*, 25–34. https://doi.org/10.1016/j.ijpe.2010.06.013 (cit. on p. 103).

Tashman, L. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437–450 (cit. on p. 155).

Teimoury, E., Modarres, M., Ghasemzadeh, F., & Fathi, M. (2010). A queueing approach to production-inventory planning for supply chain with uncertain demands: Case study of PAKSHOO Chemicals Company. *Journal of Manufacturing Systems*, *29*(2-3), 55–62. https://doi.org/10.1016/j.jmsy.2010.08.003 (cit. on pp. 66, 67).

Tempelmeier, H., & Bantel, O. (2015). Integrated optimization of safety stock and transportation capacity. *European Journal of Operational Research*, *247*(1), 101–112. https://doi.org/10.1016/j.ejor.2015.05.069 (cit. on pp. 70, 71).

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., & Murthy, R. (2009). Hive: A warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment*, *2*(2), 1626–1629 (cit. on pp. 88, 89, 109, 110, 145).

Tian, F., Willems, S., & Kempf, K. (2011). An iterative approach to item-level tactical production and inventory planning. *International Journal of Production Economics*, *133*(1, SI), 439–450. https://doi.org/10.1016/j.ijpe.2010.07.011 (cit. on pp. 70, 71).

Tinani, K., & Kandpal, D. (2017). Literature Review on Supply Uncertainty Problems: Yield Uncertainty and Supply Disruption. *Journal of the Indian Society for Probability and Statistics*, *18*(2), 89–109. https://doi.org/10.1007/s41096-017-0020-1 (cit. on p. 43).

Tiwari, S., Wee, H., & Daryanto, Y. (2018). Big data analytics in supply chain management between 2010 and 2016: Insights to industries. *Computers and Industrial Engineering*, *115*, 319–330. https://doi.org/10.1016/j.cie.2017.11.017 (cit. on p. 103).

Tookanlou, P., & Wong, H. (2019). Determining the optimal customization levels, lead times, and inventory positioning in vertical product differentiation. *International Journal of Production Economics*. https://doi.org/10.1016/j.ijpe.2019.08.014 (cit. on p. 71).

Torbacki, W., & Kijewska, K. (2019). Identifying key performance indicators to be used in logistics 4.0 and industry 4.0 for the needs of sustainable municipal logistics by means of the dematel method [3rd International Conference ”Green Cities – Green Logistics for Greener Cities”, Szczecin, 13-14 September 2018]. *Transportation Research Procedia*, *39*, 534–543. https://doi.org/https://doi.org/10.1016/j.trpro.2019.06.055 (cit. on p. 83).

Torkul, O., Yılmaz, R., Selvi, İ., & Cesur, M. (2016). A real-time inventory model to manage variance of demand for decreasing inventory holding cost. *Computers and Industrial Engineering*, *102*, 435–439. https://doi.org/10.1016/j.cie.2016.04.020 (cit. on pp. 63, 64).

Trapero, J., Cardós, M., & Kourentzes, N. (2019a). Empirical safety stock estimation based on kernel and GARCH models. *Omega (United Kingdom)*, *84*, 199–211. https://doi.org/10.1016/j.omega.2018.05.004 (cit. on pp. 44–46, 57, 58, 75).

Trapero, J., Cardós, M., & Kourentzes, N. (2019b). Quantile forecast optimal combination to enhance safety stock estimation. *International Journal of Forecasting*, *35*(1), 239–250. https://doi.org/10.1016/j.ijforecast.2018.05.009 (cit. on pp. 44, 57, 58, 75).

Trkman, P., McCormack, K., De Oliveira, M. V., & Ladeira, M. (2010). The impact of business analytics on supply chain performance. *Decision Support Systems*, *49*(3), 318–327. https://doi.org/10.1016/j.dss.2010.03.007 (cit. on pp. 2, 22, 23, 37, 103, 168).

Tsou, C. (2009). Evolutionary Pareto optimizers for continuous review stochastic inventory systems. *European Journal of Operational Research*, *195*(2), 364–371. https://doi.org/10.1016/j.ejor.2008.02.039 (cit. on p. 63).

Turban, E., Sharda, R., & Delen, D. (2011). *Decision Support and Business Intelligence Systems (9th Edition)* (P. Hall, Ed.; 9th ed.). (Cit. on pp. xiii, 19–22).

Turgut, Ö., Taube, F., & Minner, S. (2018). Data-driven retail inventory management with backroom effect. *OR Spectrum*, *40*(4), 945–968. https://doi.org/10.1007/s00291-018-0511-9 (cit. on pp. 63, 64).

Tyworth, J., & O'Neill, L. (1997). Robustness of the normal approximation of lead-time demand in a distribution setting. *Naval Research Logistics*, *44*(2), 165–186. https://doi.org/10.1002/(SICI)1520-6750(199703)44:2<165::AID-NAV2>3.0.CO;2-7 (cit. on pp. 66, 67).

Tyworth, J. (1992). Modeling transportation-inventory trade-offs in stochastic setting. *Jounal of Business Logistics*, (2), 97–127 (cit. on p. 75).

Urban, T. (2005). A periodic-review model with serially-correlated, inventory-level-dependent demand. *International Journal of Production Economics*, *95*(3), 287–295. https://doi.org/10.1016/j.ijpe.2003.11.015 (cit. on p. 63).

Uthayakumar, R., & Parvathi, P. (2011). Inventory model with pricing tactics for demand in auto-correlated products. *International Journal of Advanced Manufacturing Technology*, *52*(5-8), 833–840. https://doi.org/10.1007/s00170-010-2755-7 (cit. on p. 66).

van der Rhee, B., Schmidt, G., & Tsai, W. (2017). Hold Safety Inventory Before, At, or After the Fan-Out Point? *Production and Operations Management*, *26*(5), 817–835. https://doi.org/10.1111/poms.12676 (cit. on p. 71).

Van Donselaar, K., & Broekmeulen, R. (2013). Determination of safety stocks in a lost sales inventory system with periodic review, positive lead-time, lot-sizing and a target fill rate. *International Journal of Production Economics*, *143*(2), 440–448. https://doi.org/10.1016/j.ijpe.2011.05.020 (cit. on pp. 57, 58).

Van Kampen, T., Van Donk, D., & Van Der Zee, D. (2010). Safety stock or safety lead time: Coping with unreliability in demand and supply. *International Journal of Production Research*, *48*(24), 7463–7481. https://doi.org/10.1080/00207540903348346 (cit. on p. 43).

Vandeput, N. (2020). *Inventory optimization: Models and simulations* [cited By 0]. De Gruyter. (Cit. on pp. 45, 46, 75).

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3 (cit. on p. 53).

Vanteddu, G., Chinnam, R., Yang, K., & Gushikin, O. (2007). Supply chain focus dependent safety stock placement. *International Journal of Flexible Manufacturing Systems*, *19*(4), 463–485. https://doi.org/10.1007/s10696-008-9050-z (cit. on p. 73).

Vernimmen, B., Dullaert, W., Willemé, P., & Witlox, F. (2008). Using the inventory-theoretic framework to determine cost-minimizing supply strategies in a stochastic setting. *International Journal of Production Economics*, *115*(1), 248–259. https://doi.org/10.1016/j.ijpe.2008.05.015 (cit. on p. 61).

Vieira, A., Dias, L., Santos, M., Pereira, G., & Oliveira, J. (2019). Simulation of an automotive supply chain using big data. *Computers and Industrial Engineering*, *137*. https://doi.org/10.1016/j.cie.2019.106033 (cit. on pp. 2, 103, 104, 108–110).

Vieira, A., Dias, L., Santos, M. Y., Pereira, G., & Oliveira, J. (2020). Supply chain risk management: An interactive simulation model in a big data context [International Conference on Industry 4.0 and Smart Manufacturing (ISM 2019)]. *Procedia Manufacturing*, *42*, 140–145. https://doi.org/https://doi.org/10.1016/j.promfg.2020.02.035 (cit. on p. 86).

Voyant, C., Notton, G., Kalogirou, S., Nivet, M., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582. https://doi.org/10.1016/j.renene.2016.12.095 (cit. on p. 152).

Waller, M., & Fawcett, S. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010 (cit. on pp. 22, 103).

Wang, G., Gunasekaran, A., Ngai, E., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110. https://doi.org/10.1016/j.ijpe.2016.03.014 (cit. on pp. 19, 20, 22, 23, 103).

Wang, H., & Wang, Z. (2013). Reasearch on the real linkage effect between key parameters in inventory management. *Journal of Applied Sciences*, *13*(18), 3752–3756. https://doi.org/10.3923/jas.2013.3752.3756 (cit. on p. 65).

Wang, P., Zinn, W., & Croxton, K. (2010). Sizing inventory when lead time and demand are correlated. *Production and Operations Management*, *19*(4), 480–484. https://doi.org/10.1111/j.1937-5956.2009.01109.x (cit. on pp. 44, 56, 58).

Wang, T., & Toktay, B. (2008). Inventory management with advance demand information and flexible delivery. *Management Science*, *54*(4), 716–732. https://doi.org/10.1287/mnsc.1070.0831 (cit. on pp. 63, 64).

Wang, X., Yang, K., & Liu, T. (2019). The implementation of a practical agricultural big data system. *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, 1955–1959. https://doi.org/10.1109/ICCC47050.2019.9064475 (cit. on p. 85).

Willems, S. P. (2008). Real-world multiechelon supply chains used for inventory optimization. *Manufacturing and Service Operations Management*, *10*(1), 19–23. https://doi.org/10.1287/msom.1070.0176 (cit. on p. 72).

Winkelhaus, S., & Grosse, E. H. (2020). Logistics 4.0: A systematic review towards a new logistics system. *International Journal of Production Research*, *58*(1), 18–43 (cit. on p. 83).

Witten, I., Frank, E., Hall, M., & Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier Inc. (Cit. on pp. 115, 117, 120, 151).

Woerner, S., Laumanns, M., & Wagner, S. (2018). Joint optimisation of capacity and safety stock allocation. *International Journal of Production Research*, *56*(13), 4612–4628. https://doi.org/10.1080/00207543.2017.1380323 (cit. on p. 71).

Woerner, S., Laumanns, M., & Wagner, S. M. (2018). Simulation-Based Optimization of Capacitated Assembly Systems under Beta-Service Level Constraints. *Decision Sciences*, *49*(1), 180–217. https://doi.org/10.1111/deci.12260 (cit. on pp. 72, 73).

Wollschlaeger, M., Sauter, T., & Jasperneite, J. (2017). The Future of Industrial Communication: Automation Networks in the Era of the Internet of Things and Industry 4.0. *IEEE Industrial Electronics Magazine*, *11*(1), 17–27. https://doi.org/10.1021/ie50124a022 (cit. on p. 15).

Xiao, G., Yang, N., & Zhang, R. (2015). Dynamic pricing and inventory management under fluctuating procurement costs. *Manufacturing and Service Operations Management*, *17*(3), 321–334. https://doi.org/10.1287/msom.2015.0519 (cit. on p. 66).

Xu, H., Yu, W., Griffith, D., & Golmie, N. (2018). A Survey on Industrial Internet of Things: A Cyber-Physical Systems Perspective. *IEEE Access*, *6*, 1–1. https://doi.org/10.1109/ACCESS.2018.2884906 (cit. on pp. xiii, 15–18, 22, 103).

Xu, L. D., He, W., & Li, S. (2014). Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, *10*(4), 2233–2243. https://doi.org/10.1109/TII.2014.2300753 (cit. on pp. 16, 17).

Xu, X., Zhao, Y., & Chen, C.-Y. (2016). Project-driven supply chains: integrating safety-stock and crashing decisions for recurrent projects. *Annals of Operations Research*, *241*(1-2), 225–247. https://doi.org/10.1007/s10479-012-1240-0 (cit. on p. 73).

Yamazaki, T., Shida, K., & Kanazawa, T. (2016). An approach to establishing a method for calculating inventory. *International Journal of Production Research*, *54*(8), 2320–2331. https://doi.org/10.1080/00207543.2015.1076179 (cit. on pp. xvi, 43, 45).

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061 (cit. on pp. 118, 154).

Yang, L. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of Industrial Information Integration*, *6*, 1–10. https://www.sciencedirect.com/science/article/pii/S2452414X17300043 (cit. on pp. 15, 22, 103).

Yang, M., & Lo, M. (2011). Considering single-vendor and multiple-buyers integrated supply chain inventory model with lead time reductiong. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *225*(5), 747–759. https://doi.org/10.1243/09544054JEM1710 (cit. on pp. 63, 64).

Yao, Z., Lee, L., Jaruphongsa, W., Tan, V., & Hui, C. (2010). Multi-source facility location-allocation and inventory problem. *European Journal of Operational Research*, *207*(2), 750–762. https://doi.org/10.1016/j.ejor.2010.06.006 (cit. on p. 70).

Yavas, V., & Ozkan-Ozen, Y. D. (2020). Logistics centers in the new industrial era: A proposed framework for logistics center 4.0. *Transportation Research Part E: Logistics and Transportation Review*, *135*, 101864. https://doi.org/https://doi.org/10.1016/j.tre.2020.101864 (cit. on p. 83).

Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, *77*(2), 175–183 (cit. on p. 162).

Ye, J., Chow, J.-H., Chen, J., & Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. *Proceedings of the 18th ACM conference on Information and knowledge management*, 2061–2064 (cit. on p. 152).

You, F., & Grossmann, I. (2008). Design of responsive supply chains under demand uncertainty. *Computers and Chemical Engineering*, *32*(12), 3090–3111. https://doi.org/10.1016/j.compchemeng.2008.05.004 (cit. on pp. 57, 58).

You, F., & Grossmann, I. (2011). Balancing responsiveness and economics in process supply chain design with multi-echelon stochastic inventory. *AIChE Journal*, *57*(1), 178–192. https://doi.org/10.1002/aic.12244 (cit. on pp. 63, 70, 71).

You, F., & Grossmann, I. (2010). Integrated Multi-Echelon Supply Chain Design with Inventories Under Uncertainty: MINLP Models, Computational Strategies. *AIChE Journal*, *56*(2), 419–440. https://doi.org/10.1002/aic.12010 (cit. on p. 70).

Yu, Y., Xiong, W., & Cao, Y. (2015). A conceptual model of supply chain risk mitigation: The role of supply chain integration and organizational risk propensity. *Journal of Coastal Research*, 95–98. https://doi.org/10.2112/SI73-017.1 (cit. on p. 38).

Yue, D., & You, F. (2013). Planning and scheduling of flexible process networks under uncertainty with stochastic inventory: MINLP models and algorithm. *AIChE Journal*, *59*(5), 1511–1532. https://doi.org/10.1002/aic.13924 (cit. on pp. 70, 71).

Yue, L., Wangwei, J., Jianguo, Z., Junjun, G., Jiazhou, Z., & Aiping, J. (2016). Product life cycle based demand forecasting by using artificial bee colony algorithm optimized two-stage polynomial fitting (X. Z. Li K., Ed.). *Journal of Intelligent and Fuzzy Systems*, *31*(2), 825–836. https://doi.org/10.3233/JIFS-169014 (cit. on pp. 63, 64, 76).

Zadeh, A., Sharda, R., & Kasiri, N. (2016). Inventory record inaccuracy due to theft in production-inventory systems. *International Journal of Advanced Manufacturing Technology*, *83*(1-4), 623–631. https://doi.org/10.1007/s00170-015-7433-3 (cit. on p. 65).

Zahraei, S., & Teo, C. (2018). Optimizing a recover-and-assemble remanufacturing system with production smoothing. *International Journal of Production Economics*, *197*, 330–341. https://doi.org/10.1016/j.ijpe.2018.01.016 (cit. on pp. 57, 58).

Zhang, P., Yan, H., & Pang, K. (2019). Inventory sharing strategy for disposable medical items between two hospitals. *Sustainability (Switzerland)*, *11*(22). https://doi.org/10.3390/su11226428 (cit. on pp. 63, 64).

Zhang, Z., Li, Y., & Huang, G. (2014). An inventory-theory-based interval stochastic programming method and its application to Beijing's electric-power system planning. *International Journal of Electrical Power and Energy Systems*, *62*, 429–440. https://doi.org/10.1016/j.ijepes.2014.04.060 (cit. on p. 66).

Zhao, X., Lai, F., & Lee, T. (2001). Evaluation of safety stock methods in multilevel material requirements planning (MRP) systems. *Production Planning & Control*, *12*(8), 794–803. https://doi.org/10.1080/095372800110052511 (cit. on pp. 55, 58).

Zhong, R., Newman, S., Huang, G., & Lan, S. (2016). Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers and Industrial Engineering*, *101*, 572–591. https://doi.org/10.1016/j.cie.2016.07.013 (cit. on pp. 108, 109).

Zhou, C., & Viswanathan, S. (2011). Comparison of a new bootstrapping method with parametric approaches for safety stock determination in service parts inventory systems. *International Journal of Production Economics*, *133*(1), 481–485. https://doi.org/10.1016/j.ijpe.2010.09.021 (cit. on pp. 56, 58).

Zhou, S., & Chao, X. (2014). Dynamic pricing and inventory management with regular and expedited supplies. *Production and Operations Management*, *23*(1), 65–80. https://doi.org/10.1111/poms.12047 (cit. on p. 66).

Zipkin, P. H. (2000). *Foundations of inventory management*. McGraw-Hill. (Cit. on pp. 137, 174).

Zsidisin, G. (2003). Managerial Perceptions of Supply Risk. *Journal of Supply Chain Management*, *39*(4), 14–26. https://doi.org/10.1111/j.1745-493X.2003.tb00146.x (cit. on p. 41).