POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Conception d'un système de gestion de l'inventaire pour un portefeuille de produits à profil de demande mixte

PHILIPPE ST-AUBIN
Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie industriel

Décembre 2020

© Philippe St-Aubin, 2020.

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Conception d'un système de gestion de l'inventaire pour un portefeuille de produits à profil de demande mixte

présentée par **Philippe ST-AUBIN**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Robert PELLERIN, président
Bruno AGARD, membre et directeur de recherche
Christophe DANJOU, membre
Samira KEIVANPOUR, membre
Mustapha OUHIMMOU, membre externe

DÉDICACE

À ma mère qui a "toujours su que je ferais un doctorat",
et à Sofia, pour son support inconditionnel.

REMERCIEMENTS

Je tiens à remercier les membres du jury. Merci à Robert Pellerin d'en avoir accepté la présidence. Merci à Christophe Danjou, Samira Kevanpour et Mustapha Ouhimmou d'en être membre et merci à Benoit Ozell, représentant de la directrice des études supérieures.

Merci à MITACS pour le financement de ce projet et merci notre partenaire industriel Logistik Unicorp. Je tiens à remercier particulièrement Michel Ricard qui a cru en notre projet de recherche et Guillaume Thibault pour avoir répondu à mes questions et qui a fourni toutes les données dont je pouvais avoir besoin.

Merci à Bruno de m'avoir donné ma chance dans son laboratoire. Merci pour tes enseignements et tes conseils.

Merci à mes collègues et amis du LID sans qui mes études auraient été bien moins intéressantes.

Finalement, merci à ma famille qui m'a soutenue et vivement encouragée tout au long de mes études. Un merci particulier à Thibaut qui m'a écouté lui raconter en long et en large mes travaux de recherche.

RÉSUMÉ

Une des activités centrales de la chaîne logistique est la gestion de l'inventaire puisqu'elle implique de larges sommes d'argent investies dans du matériel, ce qui a un impact important sur la profitabilité d'une entreprise. La gestion de l'inventaire vise en particulier à déterminer quand et en quelle quantité il faut commander du matériel. La démocratisation de l'accessibilité à des outils de prévision performants, automatisés et en amélioration continue modifie les problèmes importants que les chercheurs et développeurs de systèmes de gestion de l'inventaire doivent surmonter. De plus en plus, ces problèmes se focalisent autour de l'évaluation, du suivi des performances, de la sélection des modèles de prévision et de la prise de décision. Ces sujets sont abordés dans cette thèse qui propose une approche pour la conception et le développement d'un système de gestion de l'inventaire pour un portefeuille de produits à profil de demande mixte. Les profils de demande sont une caractéristique des séries temporelle et il s'avère que les traitements et méthodes de prévisions à utiliser varient selon le profil de demande des séries. Plusieurs défis et questions demeurent sans réponse en présence de séries à profil mixtes.

Notre état de l'art a soulevé quelques-uns des principaux défis. Les métriques de précision sont souvent à la base des décisions de sélection de modèles de prévision pour la gestion de l'inventaire. Or, comment faire la sélection de modèles lorsque les métriques de performance n'atteignent pas un consensus ? Une méthode pour quantifier la précision et la fiabilité des métriques de performance des modèles de prévision est développée afin de répondre à cette question. Nos résultats ont permis de mesurer la sensibilité et la fiabilité de plusieurs métriques de performance populaires et donne quelques recommandations sur quelles métriques utiliser selon différentes circonstances.

Avec cette méthode, la sélection de modèles de prévision devient plus aisée. Cependant, il n'existe toujours pas de consensus sur le lien qui existe entre la précision des modèles de prévision et les performances associées en gestion d'inventaire. La littérature ne présente pas non plus de consensus sur l'impact de la sélection multiple et les études menées jusqu'à maintenant sont basées sur la précision des résultats plutôt que sur leur utilité en mesurant les performances en gestion d'inventaire. Des connaissances additionnelles sont requises et sont fournies par nos travaux de recherche. Nos résultats ont montré que la sélection multiple donne de meilleurs résultats que la sélection globale. Les résultats ont également montré que la sélection individuelle basée sur des métriques de précision permet d'obtenir des résultats en inventaire performants comparables à une sélection basée sur les coûts. Toutefois, le lien

entre la précision et la sélection globale demeure à éclaircir.

Finalement, la prise de décision de réapprovisionnement demeure un aspect clé dont l'optimisation peut avoir un impact significatif sur la profitabilité d'une entreprise. Or, l'impact des politiques de réapprovisionnement dynamique demeure peu étudié. On propose donc d'étudier ce facteur tout en proposant une méthodologie pour optimiser une politique dynamique basée sur le cadre conceptuel de l'apprentissage par imitation. Ce cadre conceptuel permet de tirer parti de l'apprentissage automatique, une méthode ayant connu un important succès dans divers domaines d'application, pour la résolution de problème de décision. Les résultats ont montré une amélioration importante des performances en gestion d'inventaire en utilisant une politique d'approvisionnement basée sur l'apprentissage par imitation versus une politique dynamique (s_t, Q) ou statique (s, Q) classique. Les résultats ont également montré que la méthode proposée permet de générer des politiques plus robustes aux changements de performance qu'un modèle de prévision de la demande.

La thèse dans son ensemble fournit plusieurs recommandations et méthodologies pour faire la conception d'un système de gestion de l'inventaire. Une méthodologie pour mener un tel système vers l'autonomie est également présentée. Les résultats cumulés des trois contributions ont ainsi permis d'accumuler de nouvelles connaissances sur le domaine et de proposer de nouvelles méthodes pour la résolution de problèmes d'inventaire. L'accumulation des résultats sur la relation complexe entre les performances en inventaire et la précision des prévisions de la demande a mené à l'explication de la nature de la complexité observée. Ces résultats pourront ainsi conduire au développement de nouvelles métriques de performance basées sur l'erreur et fortement corrélées avec les performances d'inventaire. Ceci permettra de faire la sélection et l'optimisation des paramètres de modèles de prévision de la demande pour la gestion d'inventaire sans avoir recours à la simulation.

Les méthodologies proposées ont été validées à l'aide de données réelles provenant de notre partenaire industriel. Les conclusions, outils et méthodes développées dans cette recherche sont en cours d'implantation chez le partenaire industriel à des fins d'utilisation en production.

ABSTRACT

One of the central activities of the logistics chain is inventory management since it involves large sums of money invested in stock, which has a significant impact on a company's profitability. Inventory management is concerned with determining when and in what quantity to order material. The democratization of the accessibility to high-performance automated and continuously improving forecasting tools is changing the important problems that researchers and developers of inventory management systems must overcome. Increasingly, these issues focus around evaluation, performance monitoring, selection of forecasting models, and decision-making. These topics are addressed in this thesis which proposes an approach for the design and development of an inventory management system for a portfolio of products with a mixed demand profile. Demand profiles are a characteristic of time series and it turns out that the treatments and forecasting methods to be used vary depending on the demand profile of the series. Several challenges and questions remain unanswered in the presence of series with mixed profiles.

Our state of the art has raised some of the main challenges. Precision metrics are often the basis for decisions on the selection of forecasting models for inventory management. However, how do you select models when performance metrics do not reach consensus? A method to quantify the precision and reliability of performance metrics of forecasting models is developed to answer this question. Our results have measured the sensitivity and reliability of several popular performance metrics and make some recommendations on which metrics to use under different circumstances.

With this method, the selection of forecast models becomes easier. However, there is still no consensus on the relation between the accuracy of forecasting models and the associated performance in inventory management. There is also no consensus in the literature on the impact of multiple selection and the studies conducted to date are based on the accuracy of the results rather than their usefulness in measuring inventory management performance. Additional knowledge is required and is provided by our research. Our results showed that multiple selection gives better results than overall selection. The results also showed that individual selection based on precision metrics achieves efficient inventory results comparable to selection based on cost. However, the relation between precision and overall selection remains to be clarified.

Finally, the decision-making of replenishment remains a key aspect whose optimization can have a significant impact on the profitability of a company. However, the impact of dynamic

replenishment policies remains little studied. We therefore propose to study this factor while proposing a methodology to optimize a dynamic policy based on the conceptual framework of learning by imitation. This conceptual framework makes it possible to take advantage of machine learning, a method that has been very successful in various application areas, for decision problems. The results showed a significant improvement in inventory management performance using an inventory policy based on imitation learning versus a dynamic (s_t, Q) or static (s, Q) policy. The results also showed that the proposed method makes it possible to generate policies that are more robust to changes in performance of a demand forecasting model.

The thesis provides several recommendations and methodologies for making the design of an inventory management system. A methodology for leading such a system towards autonomy is also presented. The cumulative results of the three contributions have thus made it possible to accumulate new knowledge in the field and to propose new methods for solving inventory problems. The accumulation of results on the complex relationship between inventory performance and the accuracy of demand forecasts has led to the explanation of the nature of the observed complexity. These results may thus lead to the development of new performance metrics based on error and strongly correlated with inventory performance. This will allow selection and optimization of demand forecasting model parameters for inventory management without resorting to simulations.

The proposed methodologies have been validated using real data from our industrial partner. The conclusions, tools and methods developed in this research are being implemented by the industrial partner for use in production.

TABLE DES MATIÈRES

DÉDICACE iii

REMERCIEMENTS iv

RÉSUMÉ v

ABSTRACT vii

TABLE DES MATIÈRES ix

LISTE DES TABLEAUX xiii

LISTE DES FIGURES xiv

LISTE DES SIGLES ET ABRÉVIATIONS xvi

CHAPITRE 1 INTRODUCTION 1

CHAPITRE 2 REVUE DE LITTÉRATURE 5
- 2.1 Introduction 5
- 2.2 Gestion de l'inventaire 5
  - 2.2.1 Politique du point de commande (r,Q) 6
  - 2.2.2 Politique de commande jusqu'au niveau (s,S) 7
  - 2.2.3 Minimiser les coûts 8
- 2.3 Prévision de la demande 9
  - 2.3.1 Demande intermittente 10
  - 2.3.2 Évaluation des prévisions 12
  - 2.3.3 Sélection de modèle 14
- 2.4 Conclusion 16

CHAPITRE 3 DÉMARCHE ET ORGANISATION 18
- 3.1 Introduction 18
- 3.2 Les objectifs de recherche 18
- 3.3 Contexte industriel 20
- 3.4 Méthodologie 22
  - 3.4.1 Système de prévision de la demande 22

3.4.2 Simulateur de la gestion d'inventaire 24
3.5 Conclusion 25

CHAPITRE 4 ARTICLE 1 : A METHODOLOGY TO EVALUATE FORECASTS PERFORMANCE METRICS 26
4.1 Introduction 26
4.2 Previous Work 27
4.3 Methodology 28
4.3.1 Build Models 29
4.3.2 Select performance Metrics 30
4.3.3 Sensitivity 31
4.3.4 Reliability 31
4.4 Results 34
4.4.1 Data 34
4.4.2 Standard deviation sensitivity 34
4.4.3 Bias sensitivity 36
4.4.4 Standard deviation-Bias sensitivity 37
4.4.4.1 Standard deviation in function of bias 37
4.4.4.2 Bias in function of standard deviation 39
4.4.5 Reliability of the metrics 41
4.4.5.1 Reliability to a change in standard deviation 42
4.4.5.1.1 Same order of magnitude 42
4.4.5.1.2 High bias 42
4.4.5.1.3 Small bias 44
4.4.5.2 Reliability to change in bias 46
4.4.5.2.1 High standard deviation and high bias 46
4.4.5.2.2 High standard deviation and low bias 46
4.4.5.2.3 Small standard deviation 46
4.5 Result Analysis 49
4.6 Conclusion 51

CHAPITRE 5 ARTICLE 2: MULTIPLE FORECASTING MODELS SELECTION IN THE CONTEXT OF INVENTORY MANAGEMENT 53
5.1 Introduction 53
5.2 Previous Work 54
5.3 Methodology 55
5.3.1 Data Partition 56

5.3.2 Forecast Demand 57
5.3.3 Forecast Accuracy 57
5.3.4 Calculate Inventory Policy Parameters 57
5.3.5 Inventory simulation 58
5.3.6 Inventory performance 59
5.3.7 Model selection 59
5.3.8 Cross-validation 60
5.4 Experiment 61
5.4.1 Benchmarks and contribution 62
5.4.1.1 Confirm the validity of the methodology 62
5.4.1.2 Estimate Performance Lift 63
5.5 Results 64
5.5.1 Simulation versus Accuracy selection 66
5.5.2 Multi versus single selection 67
5.5.3 Impact of cost on lift 67
5.6 Conclusion 70

CHAPITRE 6 ARTICLE 3: AN IMITATION LEARNING APPROACH TO INVENTORY MANAGEMENT 71
6.1 Introduction 71
6.2 Previous works 72
6.3 Methodology 73
6.3.1 Learning the Inventory Policy 74
6.3.1.1 Simulation 74
6.3.1.1.1 Forecast Demand 75
6.3.1.1.2 Base Heuristic 76
6.3.1.1.3 Extract Features 76
6.3.1.2 Set Label (optimal policy) 76
6.3.1.3 Machine learning 78
6.3.2 Validation 79
6.3.2.1 Validation Simulation 79
6.3.2.2 Performance 79
6.4 Experiment and contribution 79
6.4.1 Data 80
6.4.2 Simulation 80
6.4.3 Machine Learning 81

6.4.4 Benchmarks 81
6.5 Results 82
6.5.1 Forecasting Models Performance 82
6.5.2 Inventory Policies Performance 83
6.6 Conclusion 86

CHAPITRE 7 DISCUSSION GÉNÉRALE 88

CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS 92

RÉFÉRENCES 95

LISTE DES TABLEAUX

Tableau 2.1 Classification, description et critique des métriques de performance . 13
Table 4.1 Configurations of standard deviation, bias and variation standard deviation and bias to test and evaluate the reliability of metrics 32
Table 4.2 Average variation rate of performance metrics for a variation in standard deviation and fixed bias of forecast models 34
Table 4.3 Average variation rate of performance metrics for a variation in bias and fixed standard deviation of forecast models 37
Table 4.4 Values of each parameter for all of the different configurations 41
Table 4.5 Best choice of metric for different cases based on the reliability results 44
Table 5.1 Experiment 2: SMSS and SMRS results 71
Table 5.2 Experiment 3: SMSS results 71
Table 5.3 Configurations global performance 72
Table 5.4 Mean lift according to selection method 73
Table 5.5 Configurations' global performance with realistic costs 75
Table 5.6 Mean lift according to selection method with realistic costs 76
Table 6.1 Forecasting models accuracy (year 2017 out-of-sample) 89
Table 6.2 Forecasting models accuracy (year 2018 out-of-sample) 89
Table 6.3 Inventory Performance 91
Table 6.4 Average Performance Lift 92
Tableau 7.1 Précisions des modèles de prévision 97

LISTE DES FIGURES

Figure 2.1 Classification des séries d'après la classification de (Syntetos et al., 2005) et de (Kostenko and Hyndman, 2006) 15
Figure 3.1 Méthodologie 19
Figure 3.2 Structure de la chaîne Logistique 21
Figure 3.3 Classification des SKU d'après la classification de (Syntetos et al., 2005) 21
Figure 3.4 Entrants, extrants et paramètres des systèmes implémentés 23
Figure 4.1 Experiment diagram 29
Figure 4.2 Representation of the averaging variables to estimate performance . . 33
Figure 4.3 Standard deviation sensitivity of scaled performance metrics in absence of bias 35
Figure 4.4 Bias sensitivity of scaled performance metrics with 1% standard deviation in forecast models 36
Figure 4.5 Standard deviation variation rate in function of bias 39
Figure 4.6 Bias variation rate in function of standard deviation 41
Figure 4.7 Spearman Rank correlation and nDCG for $(+++)$ configuration . . 46
Figure 4.8 Spearman Rank correlation and nDCG for $(---)$ configuration . . 47
Figure 4.9 Spearman Rank correlation and nDCG for $(++-)$ configuration . . 48
Figure 4.10 Spearman Rank correlation and nDCG for $(-+-)$ configuration . . 49
Figure 4.11 Spearman Rank correlation and nDCG for $(+-+)$ configuration . . 50
Figure 4.12 Spearman Rank correlation and nDCG for $(+--)$ configuration . . 51
Figure 4.13 Spearman Rank correlation and nDCG for $(+++)$ configuration . . 52
Figure 4.14 Spearman Rank correlation and nDCG for $(++-)$ configuration . . 53
Figure 4.15 Spearman Rank correlation and nDCG for $(+--)$ configuration . . 54
Figure 4.16 Spearman Rank correlation and nDCG for $(-++)$ configuration . . 55
Figure 4.17 Spearman Rank correlation and nDCG for $(-+-)$ configuration . . 56
Figure 4.18 Spearman Rank correlation and nDCG for $(---)$ configuration . . 57
Figure 4.19 Standard deviation sensitivity when averaging results of 1, 10 and 100 series 58
Figure 4.20 Reliability of series in function of the level of intermittence 59
Figure 5.1 Simulation flow chart 63
Figure 5.2 Data Partition 64
Figure 5.3 Experiment 1: Multi-Model Selection 68
Figure 5.4 Experiment 2: Classical Accuracy Selection 70

Figure 5.5 Experiment 3: Classical Simulation Selection 70
Figure 5.6 All Configurations level of SO compared to the average CSL 72
Figure 5.7 Distribution of item costs 74
Figure 5.8 All Configurations TC compared to the average CSL with realistic costs 76
Figure 6.1 Imitation Learning and performance evaluation methodology 89
Figure 6.2 Simulation flow chart 90
Figure 6.3 Items cost distribution 92
Figure 6.4 Policies mean and standard deviation across all FM and SBA Client Service Level and Stock Ordered 93
Figure 6.5 Policies mean and standard deviation across all FM and SBA Client Service Level and Total Cost 94
Figure 7.1 Exemple de prévision pour une série intermittente 97
Figure 7.2 Quantité en stock à chaque période selon les recommandations du modèle 98

LISTE DES SIGLES ET ABRÉVIATIONS

CV	Cross Validation
EOQ	Economic Order Quantity
FM	Forecasting Model
IL	Imitation Learning
IP	Inventory Policy
MASE	Mean Absolute Scaled Error
ML	Machine Learning
MSE	Mean Squared Error
PIS	Period in Stock
RSL	Realized Service Level
sAPIS	scaled Absolute Period in Stock
SBA	Syntetos and Boylan Approximation
SES	Lissage Exponentiel Simple
SL	Service Level
SMA	Moyenne Mobile
sRMSE	scaled Root Mean Squared Error
ss	Safety Stock
TSL	Target Service Level

CHAPITRE 1 INTRODUCTION

Il existe au sein d'une même entreprise plusieurs objectifs concurrents. Par exemple, le département de marketing souhaite habituellement maximiser les revenus et par conséquent conserver un stock en inventaire élevé afin de répondre à la demande des clients. Le département responsable de la production souhaite minimiser les coûts de production et donc favorisera de longs cycles de production avec un minimum d'arrêts. Cela requiert de conserver un stock de matière première et d'en-cours élevés. Le département des finances souhaiterait minimiser les investissements pour disposer d'un maximum de capital ce qui se traduit par un minimum de stocks de matière première et de produits finis (Arnold, 2008).

C'est entre autres raisons le besoin d'un arbitrage entre les objectifs concurrents des entreprises qui a mené au regroupement des activités en lien avec les flux de matière au sein d'une même fonction : la gestion de la chaîne logistique. Une des activités centrales de la chaîne logistique est la gestion de l'inventaire puisqu'elle implique de larges sommes d'argent investies dans du matériel ce qui a un impact important sur la profitabilité d'une entreprise (Chen et al., 2005). Cette activité vise donc à déterminer quand et en quelle quantité il faut commander du matériel. Étant donné son impact sur l'ensemble des activités des entreprises, elle offre dans plusieurs cas un important potentiel d'économie (Axsäter, 2006). Il existe plusieurs méthodes pour faire la gestion de l'inventaire et de plus en plus ces méthodes reposent sur la prévision de la demande (Prak and Teunter, 2019).

La prévision de la demande contient un grand ensemble de modèles et de méthodes de prévision spécialisés pour différents types de séries. Depuis l'approche proposée par Croston (1972), une nouvelle classe de série temporelle a été identifiée : les séries temporelles intermittentes. Ce type de profil pour les séries temporelles se caractérise par la présence de périodes consécutives d'amplitude nulle ou par une grande variation de l'amplitude de la série. Il a été démontré que le profil de la demande d'un produit constitue un facteur ayant un impact important sur la précision des prévisions (Teunter and Duncan, 2009) et sur les performances en gestion de l'inventaire (Sani and Kingsman, 1997). C'est pourquoi plusieurs modèles de prévision et recommandations pour la gestion de l'inventaire ont été proposés pour gérer explicitement un tel type de demande.

Bien que les recommandations et méthodes soient distinctes en fonction du type de demande, déterminer à quel seuil appliquer l'une ou l'autre des recommandations n'est pas évident. Syntetos et al. (2005) ont proposé des seuils pour permettre de caractériser et de quantifier théoriquement les propriétés des séries qui permettent de savoir quelle méthode devrait

s'appliquer. Adaptés ensuite par Kostenko and Hyndman (2006), ces seuils ne semblent pas permettre en pratique de prédire si un modèle de prévision performera mieux qu'un autre sur une série (Kourentzes, 2014) ; (do Rego and de Mesquita, 2015).

Cela révèle un besoin à combler au niveau de la sélection des modèles de prévision spécialement en contexte ou des séries de tout type de profils sont présentes. De plus, ce projet de recherche s'insère dans un contexte où les outils de prévision entièrement automatisés disponibles en open source, intégrés aux systèmes de gestion d'entreprise ou sur des plateformes de services web, se démocratisent et sont en amélioration constante. Cela implique que, les problèmes importants qu'ont à résoudre les chercheurs et développeurs de systèmes de gestion d'inventaire sont de plus en plus focalisés autour de l'évaluation, du suivi des performances, de la sélection des modèles de prévisions et de la prise de décision.

Pour faire la sélection de modèles, en général quelques métriques de performance sont utilisées. La recherche sur les métriques de performance pour les modèles de prévision s'est essentiellement concentrée autour de la conception de métriques applicables pour tous les types de séries et permettant la comparaison des performances entre séries d'échelles différentes (Hyndman and Koehler, 2006), (Armstrong and Collopy, 1992). Il s'en est suivi le développement de plusieurs métriques d'erreur, sans toutefois que des directives claires pour savoir à quelle métrique accorder le plus de confiance au moment de la sélection ne suivent. Par exemple, en cas de désaccord entre les rangs attribués à différents modèles de prévision par différentes métriques de performance, à quelle métrique devrait-on accorder le plus de confiance pour déterminer le meilleur modèle ?

De plus en plus, la prévision de la demande et la décision de réapprovisionnement sont étudiées conjointement alors que ces deux sujets ont traditionnellement été considérés comme distincts (Prak and Teunter, 2019). Ainsi, pour garantir l'utilité des prévisions, Gardner (1990) eut en premier l'idée de les évaluer dans les conditions dans lesquels elles seraient utilisées. Depuis, plusieurs tentent de déterminer la relation qui existe entre la précision des prévisions, évaluée par des métriques de performance basées sur l'erreur, et leur utilité dans leur contexte d'utilisation (en gestion de l'inventaire). Les résultats demeurent à ce jour mitigé. En effet, certains ne trouvent pas de lien direct entre la précision et l'utilité (Teunter and Duncan, 2009), (Kourentzes, 2013) alors que d'autres voient une corrélation positive (Sanders and Graman, 2009), (Syntetos et al., 2010a). Un facteur qui semble être significatif sur la présence ou non de corrélation est le profil de demande des produits. Apparemment, en présence de profils intermittents, l'utilité n'est pas corrélée avec la précision.

Toutes ces considérations mises bout à bout révèlent plusieurs questions et difficultés rencontrées dans la conception d'un système de gestion de l'inventaire. Ce projet de recherche

visera donc la conception et l'implantation d'un système de gestion de l'inventaire pour un ensemble de produits à profil de demande varié. Ce faisant, certaines questions et difficultés actuellement rencontrées dans l'état de l'art seront abordées à travers trois objectifs spécifiques.

Tel que mentionné, plusieurs métriques de performances basées sur l'erreur sont utilisées en général pour faire la sélection de modèles de prévision. En cas de résultats mitigés, comment savoir à laquelle accorder le plus d'importance ? Cette question revient à se demander comment évaluer les métriques de performance. Ce sujet sera abordé dans cette thèse au chapitre 4 où une méthode pour quantifier la sensibilité et la fiabilité des métriques de performance basées sur l'erreur est proposée. L'idée principale pour cet objectif est de simuler les résultats obtenus d'un grand ensemble de modèles dont les paramètres de la distribution d'erreur sont connus.

Le lien qui existe entre la précision des méthodes de prévision et les performances associées en contexte de gestion d'inventaire reste à démontrer. De plus, les approches de classification théoriques des profils ne semblent pas en mesure d'identifier quel modèle appliquer à quel type de série. Une méthode de sélection de modèles de prévision en contexte de gestion d'inventaire est donc requise pour en étudier les résultats et apporter des conclusions sur l'impact de la sélection multiple en contexte de gestion d'inventaire. Ceci sera abordé au chapitre 5.

Une fois les modèles de prévisions choisis, une décision de réapprovisionnement est requise pour assurer l'approvisionnement de l'inventaire. Puisque la demande pour un produit peut changer rapidement et ainsi les prévisions de la demande devenir moins adéquates, une politique d'inventaire doit être en mesure de s'adapter aux changements de performance des modèles de prévision. Le chapitre 6 propose donc de mesurer ce facteur sur une politique statique (s, Q) et dynamique (s_t, Q) tout en proposant une méthode pour entraîner une politique de réapprovisionnement dynamique basée sur le cadre conceptuel de l'apprentissage par imitation.

La résolution de l'objectif principal servira à répondre à une problématique industrielle réelle d'une entreprise du secteur de la distribution, Logistik Unicorp (LU), notre partenaire industriel qui offre un service de gestion de l'uniforme. LU reçoit chaque jour la demande des membres des organisations clientes via une application web. La demande de chaque membre individuel pour chaque item de l'uniforme est ainsi enregistrée. Chacun des objectifs spécifiques est implémenté et validé sur ce cas d'étude réel. Les recommandations et méthodes développées le sont dans l'objectif d'être implantées en production au sein de l'entreprise partenaire.

La suite de cette thèse est organisée comme suit : le chapitre 2 propose une revue de littérature

qui situera le problème dans son cadre théorique et présentera les principales familles d'outils utilisées pour la résolution des problèmes. Le chapitre 3 présentera les détails et particularités du problème, de son cas d'étude et de notre approche de résolution. Les chapitres 4 à 6 présentent les contributions méthodologiques et scientifiques de la thèse. Ces trois chapitres ont été soumis à des journaux pour évaluation par les pairs et publication. Ces chapitres sont suivis par une discussion des résultats au chapitre 7. Les conclusions et perspectives de recherche sont présentées au chapitre 8.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Introduction

Le chapitre précédent a introduit l'objectif général, visant la conception et le développement d'un système de gestion de l'inventaire pour un portefeuille de produits à profil de demande mixte.

Ce chapitre se consacrera donc à situer le problème au sein du domaine de la gestion de l'inventaire à la section 2.2, puis présentera les principaux outils utilisés pour le résoudre à la section 2.3. La section 2.4 présente finalement une conclusion critique des travaux antérieurs.

2.2 Gestion de l'inventaire

La gestion de l'inventaire est un sujet vaste qui touche plusieurs aspects d'une entreprise. On pourrait donc l'étudier sous plusieurs angles. Étant donné le contexte présenté et le cas d'étude, on limitera la revue de littérature aux travaux touchant la décision de réapprovisionnement au niveau opérationnel tel que défini par Axsäter (2006). Il définit un système de gestion de l'inventaire comme un système visant à déterminer la quantité et la période à laquelle commander du matériel.

Waters (2008) distingue les méthodes pour contrôler le stock selon le type de demande. D'une part, il y a les méthodes basées sur la demande dépendante ou multiéchelons et d'autre part, les méthodes basées sur la demande indépendante ou à échelon unique. Dans le premier cas, on retrouve les techniques de Material Requirement Planning et les méthodes de Juste-à-temps qui permettent de modéliser la dépendance entre la demande pour différents items. Le second cas considère que la demande pour chaque item est indépendante de celle des autres items. On y considère que la demande pour un item provient de l'ensemble des demandes de chaque client de manière individuelle. C'est dans ce dernier cas que se situe le projet de recherche actuel. La revue de littérature se concentrera donc à en expliquer les principales méthodes.

Au sein des méthodes pour la demande indépendante (ou à échelon unique), on distingue les politiques d'inventaire selon quelles informations sont disponibles pour la prise de décision. Ainsi, on dit des politiques d'inventaire qu'elles sont basées sur de l'information préalable si des prévisions de la demande sont disponibles. On considère les politiques comme classiques si elles se basent uniquement sur l'épuisement des stocks (Babaï and Dallery, 2005).

Les méthodes classiques contiennent principalement deux Politiques d'Inventaire (IP) : la politique du point de commande (r,Q) (Wilson, 1934) et la politique de commande jusqu'au niveau (s,S) (Clark and Scarf, 1960) ; (Karlin, 1960). Il existe également divers variantes qui sont des cas particuliers ou des hybridations de celles-ci. Ainsi que la méthode de commande économique (EOQ) (Harris, 1913) ; (Wilson, 1934) qui détermine la quantité à commander qui minimise les coûts.

Les politiques utilisées dépendent également du type de suivi de l'inventaire qui est soit périodique soit continu. Dans le cas périodique, les quantités sont révisées à intervalle fixe. Les quantités en inventaire sont donc inconnues entre les révisions. Un suivi continu permet de connaître le niveau des stocks en tout temps. Le problème de gestion de l'inventaire et le cas d'étude dans cette thèse sont ceux d'un système de gestion de la demande indépendante ou à échelon unique, avec un suivi continu et de l'information préalable sous la forme de prévision de la demande.

2.2.1 Politique du point de commande (r,Q)

La politique (r,Q) consiste à passer une commande de quantité fixe Q lorsque le stock atteint ou descend en dessous du seuil de commande r. Avec l'approche classique, déterminer le point de commande requiert une hypothèse sur la distribution de la demande. Celui-ci correspond à la demande moyenne durant la période de couverture plus un stock de sécurité (ss) : $r = \mu + ss$. Où μ correspond à la demande moyenne durant le délai d'approvisionnement et la période de couverture, un nombre de périodes durant lesquelles la commande doit pouvoir subvenir à la demande. Dans le cas d'une demande normale, le stock de sécurité (ss) correspond à l'écart-type de la demande durant la période de couverture fois un facteur de sécurité qui correspond au quantile correspondant au niveau de service cible (TSL). La quantité à commander Q peut être déterminée par la méthode de Silver-Meal (Silver, 1973) ou par EOQ.

Pour déterminer les paramètres à partir de prévisions de la demande, on peut remplacer la demande moyenne durant le délai d'approvisionnement par la somme des prévisions (F_t) durant ladite période : $r_t = \sum_{i=t+1}^{t+1+LT} F_i + ss$. Où, LT représente le délai d'approvisionnement et t la période d'évaluation du point de recommande. Cette méthode de calcul rend le paramètre r de la politique dynamique puisqu'il est réévalué chaque période (Babai et al., 2009), c'est pourquoi on note r_t.

Il existe plusieurs propositions pour estimer le stock de sécurité. Il est habituellement exprimé sous la forme $ss = k \cdot \sigma_{LT}$. Où, k est le facteur de sécurité et σ_{LT} l'écart type des erreurs de prévisions. Silver et al. (1998) et Axsäter (2006) proposent d'estimer σ_1 par $\sqrt{MSE}$, la racine de l'erreur moyenne au carré (MSE). σ_{LT} est alors estimé par $\sigma_{LT} = \sqrt{LT} \cdot \sigma_1$. D'autres

adoptent une approche algorithmique au problème de détermination du stock de sécurité ou des paramètres de la politique. Par exemple, Kim et al. (2005) a proposé une approche par renforcement (action-value) pour choisir le facteur de sécurité d'après le TSL observé. Grewal et al. (2010) et Solis (2015) ont estimé le meilleur point de recommande et le meilleur facteur de sécurité par simulation.

Le facteur Q est en général déterminé pour couvrir une quantité de périodes donnée à partir des prévisions : $Q = \sum_{i=t+1+LT}^{t+1+LT+R} F_i + r - IO$ où IO est le niveau du stock comprenant les commandes ouvertes et R est la période de couverture.

2.2.2 Politique de commande jusqu'au niveau (s,S)

La politique (s,S) consiste à passer une commande lorsque les stocks atteignent un niveau inférieur ou égal à s (le seuil de commande). La commande est passée de manière à recompléter les stocks jusqu'au niveau S. Cette politique diffère de la précédente par le fait que la quantité à commander varie en fonction du point auquel la commande est passée.

Dans le cas classique, différentes heuristiques pour la sélection des paramètres ont été proposées. Parmi celles-ci se trouvent l'heuristique de Naddor (Naddor, 1975), l'approximation de puissance (Ehrhardt, 1979) ainsi que l'approximation normale (Wagner, 1970). Dans le cas où des prévisions de la demande sont disponibles, cette politique se retrouve équivalente à la politique (r,Q).

Un facteur qui demeure peu étudié pour la détermination des paramètres des politiques de réapprovisionnement est l'impact d'une estimation dynamique des paramètres (do Rego and de Mesquita, 2015). Dans la plupart des recherches sur le sujet toutefois, l'impact d'évaluer les paramètres de façon dynamique semble positif. Par exemple, Babai and Dallery (2009) ont montré en comparant une politique (r,Q) statique et dynamique l'impact positif d'une évaluation dynamique. De même, Kanet et al. (2010) ont également montré qu'il semblait bénéfique d'évaluer le stock de sécurité dynamiquement en présence de demandes non stationnaires. Finalement, Grewal et al. (2015) ont montré sur des données simulées qu'adapter les paramètres selon le cycle des saisonnalités permet d'améliorer les résultats. De plus, il est pratique courante lorsque les paramètres des politiques sont calculés avec des prévisions de calculer les facteurs dynamiquement. Par exemple, Tiacci and Saetta (2009), Babai et al. (2009) et Syntetos et al. (2010b) ont les trois calculés les paramètres de leur politique d'approvisionnement en les réévaluant chaque période pour intégrer la donnée la plus récente.

Dans notre recherche, nous mesurons l'impact de l'évaluation dynamique des paramètres d'une politique d'inventaire. Les interactions avec les modèles de prévision de la demande sont

également considérées afin d'évaluer la robustesse des politiques dynamiques et statiques sur des variations de performances dans les prévisions. Une approche pour calculer une politique basée sur le cadre conceptuel de l'apprentissage pas imitation est également proposée. Ceci sera développé au chapitre 6, où un état de l'art complémentaire sur l'apprentissage par imitation figure.

2.2.3 Minimiser les coûts

Une alternative aux politiques d'inventaire qui visent à garantir un TSL tout en conservant un minimum de stock est de représenter le problème de gestion de l'inventaire comme un problème de minimisation des coûts. En général, des coûts de détention du stock et des coûts de rupture de stock (Syntetos et al., 2010b). Beaucoup de travaux utilisant différents outils et suivant différents courants scientifiques ont été employés pour la résolution du problème de gestion de l'inventaire.

Par exemple, plusieurs modélisent le problème de la gestion du stock à partir des coûts et des contraintes d'un problème particulier et utilisent diverses techniques d'optimisation pour minimiser les coûts. Notamment, Mohammaditabar et al. (2012) a utilisé une résolution par recuit simulé d'un modèle qui classifiait les items en groupes et déterminait simultanément une politique de gestion des stocks à appliquer sur chaque groupe pour minimiser les coûts. Kırcı et al. (2019) ont modélisé le problème de gestion de l'inventaire dans un contexte d'articles périssables comme un jeu de Stackelberg entre le détaillant et le fabricant qui doivent s'adapter à la stratégie de l'autre et où le détaillant choisit lui-même son cycle de réapprovisionnement. Movahed and Zhang (2015) ont proposé un modèle robuste de programmation linéaire en nombres entiers mixtes (MILP) pour déterminer les paramètres d'une politique (s,S) qui minimise l'espérance des coûts. Coelho and Laporte (2014) ont modélisé le problème de la livraison conjointement avec le problème de gestion de l'inventaire pour des produits périssables. Ghalebsaz-Jeddi et al. (2004) ont modélisé le problème comme un problème d'optimisation non linéaire, intégré des contraintes de budget et considéré une demande normale indépendante et identiquement distribuée.

Donc en général, les auteurs qui utilisent cette approche le font pour intégrer des contraintes additionnelles qu'il serait difficile de prendre en compte par l'utilisation de politiques d'inventaire. Une faiblesse de l'approche d'optimisation directe de la prise de décision est que l'optimisation est faite à partir de données historiques ou sur des distributions de demande connues. Alors que rien ne garantit que les paramètres demeurent optimaux une fois le modèle en production (Sani and Kingsman, 1997). Le modèle devrait donc être optimisé sur un ensemble d'entraînement et les résultats mesurés sur un ensemble de validation pour montrer

le gain réel apporté.

Une autre méthode de résolution est d'approcher le problème comme un problème de contrôle optimal ou d'apprentissage par renforcement. Dans ce cas, à chaque période le système doit prendre la décision de réapprovisionnement qui minimise les coûts. Par exemple, Van Roy et al. (1997) ont utilisé la programmation neuro dynamique pour optimiser une politique d'inventaire. Cela consiste essentiellement à utiliser de la programmation dynamique où les coûts sont estimés par un modèle paramétrique dont les paramètres sont ajustés après chaque observation des coûts réels par descente de gradient. Kara and Dogan (2018) ont utilisé du Q-learning avec un algorithme state-action-reward-state-action (sarsa) semblable à la méthode précédente, mais où cette fois les coûts d'une décision sont évalués par les valeurs Q pour entraîner une politique de réapprovisionnement. Un des désavantages de ces deux méthodes toutefois est qu'elles sont difficiles à mettre à l'échelle pour des problèmes avec beaucoup d'items. Sinon, d'autres ont approché le problème comme un problème de décision markovien (Giannoccaro and Pontrandolfo, 2002), (Zhang and Wang, 2017).

Ces méthodes sont efficaces pour minimiser les coûts. Cependant, les coûts de stockage et les coûts de rupture de stock sont particulièrement difficiles à estimer (Çetinkaya and Parlar, 1998). C'est pourquoi souvent, en pratique, des méthodes basées sur le niveau de service sont favorisées. Par exemple, dans le cas d'étude de cette thèse, le coût de rupture de stock est difficile à évaluer, car il ne revient pas à une vente perdue. Dans notre cas, si le TSL n'est pas respecté cela pourrait se traduire par la perte d'un contrat de service avec une des organisations membres. Ce qui se traduit par des pertes importantes de revenus.

2.3 Prévision de la demande

La prévision de la demande en contexte de gestion d'inventaire s'appuie sur l'estimation la demande moyenne sur un horizon de périodes (Axsäter, 2006). Cette information est utile pour la planification et la prise de décision de réapprovisionnement, car la plupart du temps la réception des items requiert un délai de livraison. Pour cette raison, les prévisions de la demande peuvent être utilisées en gestion de l'inventaire. Les méthodes les plus populaires sont présentées ici.

On compte parmi les modèles les plus couramment utilisés dans le domaine de la gestion de l'inventaire, les modèles de lissage. Ceux-ci sont une adaptation de la moyenne mobile. Le plus simple d'entre eux est le modèle de lissage exponentiel simple (SES) (Brown, 1959), où l'on peut pondérer à l'aide d'un facteur de lissage (α) compris entre 0 et 1 l'importance des

nouvelles observations par rapport aux observations historiques :

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \tag{2.1}$$

Ce modèle a été adapté pour pouvoir inclure une tendance (Holt, 1957), une saisonnalité (Winters, 1960) puis une double saisonnalité (Taylor, 2003). Une autre adaptation du lissage exponentiel est la méthode theta (Assimakopoulos and Nikolopoulos, 2000), méthode célèbre pour avoir remporté le concours de prévision M3 (Makridakis and Hibon, 2000). Cette méthode a été démontrée par Hyndman and Billah (2003) comme étant équivalente au SES avec une dérive :

$$\hat{Y}_{t+h} = \alpha Y_t + (1-\alpha)\hat{Y}_t + \frac{1}{2}\hat{b}\left(h - 1 + \frac{1}{\alpha}\right) \tag{2.2}$$

$$\hat{b} = \frac{6}{n^2-1}\left(\frac{2}{n}\sum_{t=1}^{n} tY_t - (n+1)\bar{Y}\right) \tag{2.3}$$

où n est le nombre d'observations de la série Y.

Les modèles de lissage peuvent maintenant déterminer des intervalles de confiance grâce à la formulation en espace d'état proposée par Hyndman et al. (2002) qui permet d'identifier leur modèle stochastique sous-jacent. Avant cette formulation, seuls les modèles ARIMA (Box and Jenkins, 1970) permettaient de produire des intervalles de confiance étant donné que leur formulation permet de générer la distribution des observations. Le modèle ARIMA est défini ainsi :

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right)(1-B)^d Y_t = \left(1 + \sum_{i=1}^{q} \theta_i B^i\right)\epsilon_t \tag{2.4}$$

Il est composé par la combinaison d'un modèle autorégressif d'ordre p, d'une moyenne mobile d'ordre q et d'une différence d'ordre d. B représente l'opérateur différence : $B^d Y_t = Y_t - Y_{t-d}$, ϕ_i le coefficient autorégressif d'ordre i pour la série temporelle Y_t et θ_i le coefficient de moyenne mobile d'ordre i sur la série temporelle des erreurs ϵ_t.

Les modèles ARIMA peuvent également intégrer la saisonnalité en ajoutant des termes de différences d'ordre de la saisonnalité. Cependant, pour qu'un modèle ARIMA puisse bien modéliser une série, celle-ci doit être stationnaire en différence.

2.3.1 Demande intermittente

Il se trouve que les séries de demandes issues des produits à faible vélocité de demande sont rarement stationnaires en différence. Un profil de demande à faible vélocité et non stationnaire

en différence est la demande dite intermittente. Cette demande se caractérise par des périodes consécutives de demande nulle ou bien par une forte variation dans le volume de demandes (Babiloni et al., 2010). Croston (1972) fut le premier à proposer une approche pour ce type de demande. Sa méthode divise la série de demandes en deux séries : la série des volumes de demandes (Z_t) et la série d'intermittence (nombre de périodes inter demande) (X_t). Un lissage exponentiel des deux séries est effectué et le quotient des deux prévisions utilisé pour former la prévision de la demande :

$$\hat{Z}_{t+1} = \alpha Z_t + (1-\alpha)\hat{Z}_t \tag{2.5}$$

$$\hat{X}_{t+1} = \alpha X_t + (1-\alpha)\hat{X}_t \tag{2.6}$$

$$\hat{Y}_{t+1} = \hat{Z}_{t+1}/\hat{X}_{t+1} \tag{2.7}$$

On obtient ainsi la prévision de la consommation. Toutefois, cette méthode est biaisée (Syntetos and Boylan, 2001). Syntetos and Boylan (2005) ont donc proposé de multiplier les prévisions de la méthode de Croston par un coefficient pour corriger le biais :

$$\hat{Y}_{t+1} = \left(1 - \frac{\alpha}{2}\right)\frac{\hat{Z}_{t+1}}{\hat{X}_{t+1}} \tag{2.8}$$

Cette correction est connue sous le nom de l'approximation de Syntetos et Boylan (SBA). Teunter et al. (2011) ont ensuite proposé la méthode TSB qui utilise différents coefficients de lissage et qui estime la probabilité de commande plutôt que le nombre d'intermittences pour pouvoir réévaluer la probabilité de commande à toutes les périodes au lieu de devoir attendre une nouvelle demande non nulle pour le faire :

$$\hat{p}_{t+1} = \hat{p}_t + \beta(p_t - \hat{p}_t) \tag{2.9}$$

$$\hat{Z}_{t+1} = \hat{Z}_t + \alpha p_t(Z_t - \hat{Z}_t) \tag{2.10}$$

$$\hat{Y}_{t+1} = \hat{p}_{t+1}\hat{Z}_{t+1} \tag{2.11}$$

Où, $p_t = 1$ quand une demande occure et $p_t = 0$ en absence de demande. On remarque que l'amplitude de la demande n'est pas mise à jour si la demande est nulle : $\hat{Z}_{t+1} = \hat{Z}_t$ dans le cas où $p_t = 0$. Ce modèle est utile pour modéliser le cycle de vie des items.

Malgré les promesses de meilleures performances des nouveaux modèles développés explicitement pour la demande intermittente, Teunter and Duncan (2009) ont montré que prévoir systématiquement des valeurs nulles pouvait être plus précis que d'utiliser ces méthodes. Ce qui a suscité de l'intérêt pour la recherche de nouvelles méthodes d'évaluation capables de

prendre en compte la demande intermittente. Ces méthodes sont présentées dans la section suivante.

2.3.2 Évaluation des prévisions

L'évaluation des prévisions est classiquement mesurée par la précision de celles-ci. L'évaluation de la précision repose essentiellement sur des métriques évaluant l'erreur de prévision. L'erreur de prévision est définie comme : $e_h = Y_h - F_h$, la différence entre la valeur actuelle de la période h Y_h et la prévision pour la période h, F_h.

Dans un sondage auprès de chercheurs et de praticiens Carbone and Armstrong (1982) ont identifié que les raisons qui motivaient la sélection de l'une ou l'autre des métriques de performance étaient reliées à leur facilité d'interprétation, leur facilité d'implémentation et leur rapidité d'exécution. Depuis l'arrivée des concours de prévision M (Makridakis and Hibon, 2000), les méthodes d'évaluation ont été mieux uniformisées à travers les publications. Toutefois, Billah et al. (2006) ont soulevé, malgré cela, l'absence de théorie pour décider quelle métrique utiliser. Depuis, plusieurs problèmes rencontrés avec les métriques d'évaluation ont été résolus. Notamment, les difficultés de mise à l'échelle pour une évaluation agrégée des résultats de séries à différentes échelles ou, les cas intermittents qui impliquaient souvent des divisions par zéro ou des divisions par des nombres proche de zéro. C'est en partie grâce aux propositions de Hyndman and Koehler (2006) que ces problèmes sont aujourd'hui résolus. Hyndman and Koehler (2006) divisent les métriques de performance en quatre types : Les métriques dépendantes d'échelle, basées sur le pourcentage d'erreur, relatives et mises à l'échelle. Wallström and Segerstedt (2010) proposent un cinquième type de métrique : les métriques cumulatives. Une description et critique de chaque type de métrique est présentée à la table 2.1.

Tableau 2.1 Classification, description et critique des métriques de performance

Type	Exemple	Description	Critique
Dépendante d'échelle	$MSE = mean(e_h^2)$ $MAE = mean(\|e_h\|)$ $RMSE = \sqrt{MSE}$	Métriques dont l'échelle dépend de celle de la série temporelle évaluée.	Ne permets pas de comparer les performances entre les séries.
Basée sur le pourcentage	$MAPE = mean(\|\frac{e_h}{Y_h}\|)$ $sMAPE = 2\frac{\|e_h\|}{Y_h+F_h}$	Métrique dont l'erreur est évaluée en proportion de l'amplitude de la série.	Ces métriques sont indéfinies en présence d'amplitude nulle. Pénalisent davantage les erreurs positives que négatives.
Relative	$MRAE = mean(\|\frac{e_h}{e_h^{bench}}\|)$ $RelMAE = \frac{MAE}{MAE^{bench}}$	Métrique dont l'erreur ou la métrique est relative à celle d'une méthode de référence	Moyenne indéfinie et variance infinie.
Mise à l'échelle	$MASE = mean\left(\frac{\|e_h\|}{MAE^{bench}_{in-sample}}\right)$	Métrique dont l'erreur est normalisée par un facteur in-sample	Peut être biaisée vers 0.
Cumulative	$PIS = \sum_{h=1}^{H} -e_h$	Métrique dont les erreurs sont cummulées	Ne permets pas de comparer les performances entre les séries.

Dans la table 2.1, l'indice *bench* signifie que la valeur a été calculée à partir d'une méthode de référence, habituellement une méthode naïve. L'indice $in - sample$ signifie que la valeur est calculée sur les données d'entraînement.

Avec les métriques mises à l'échelle (Hyndman and Koehler, 2006) et les métriques cumulatives (Wallström and Segerstedt, 2010), il devient possible d'évaluer la précision des modèles de prévisions même dans des cas intermittents et d'agréger les résultats à travers un ensemble de séries d'échelles différentes sans tomber sur des valeurs indéfinies ou infinies. Les métriques non cumulatives peuvent toutefois demeurées biaisées favorablement vers des prévisions nulles dans le cas intermittent et tel qu'observé par (Teunter and Duncan, 2009).

Bien que l'on sache aujourd'hui comment évaluer les performances agrégées d'un ensemble de séries à profil mixte, il n'en demeure pas moins qu'il n'existe toujours aucune théorie pour choisir la "meilleure" métrique. En général, plusieurs métriques sont utilisées pour évaluer les performances d'un modèle de prévision. Mais que faire si les résultats des métriques sont conflictuels ? Très peu d'études ont abordé ce sujet. Les études menées pour comparer et guider le choix des métriques se sont surtout attardées sur les propriétés mathématiques de celles-ci (Hyndman and Koehler, 2006) ou encore sur des comparaisons empiriques et statistiques entre les résultats des différentes métriques (Armstrong and Collopy, 1992), (Wallström and Segerstedt, 2010). Aucune de ces études n'a tenté de quantifier la précision des métriques pourtant énoncée comme un facteur important par Armstrong and Collopy (1992). De plus, comment choisir à laquelle faire le plus confiance en cas de résultat conflictuel ? Les études précédemment citées nous indiquent quelles métriques sont corrélées sans donner d'indication sur la fiabilité et donc sur l'importance à accorder à chacune d'elles. Une piste de solution à ce problème est proposée dans nos recherches au chapitre 4.

2.3.3 Sélection de modèle

Les métriques de performance basées sur l'erreur (métriques d'erreur) servent d'indicateur pour guider la sélection de modèles de prévision. Les compétitions M (Makridakis and Hibon, 2000) (Makridakis et al., 2018) ont d'ailleurs permis de guider la communauté des prévisionnistes et des chercheurs en procurant et en raffinant une méthodologie d'évaluation des modèles de prévisions sur un grand ensemble de séries temporelles avec profils de demande mixtes, agrégation temporelle variée et de nature diverses (logistique, économique, etc.). Il est généralement accepté que le "meilleur" modèle est celui qui obtient la meilleure précision (évaluée avec les métriques d'erreur) sur l'ensemble des séries.

D'autres recherches tentent plutôt d'identifier des facteurs théoriques de classification des séries qui permettraient de sélectionner le modèle de prévision le plus adapté sans nécés-

siter d'évaluation empirique. (Syntetos et al., 2005) ont proposé une telle classification qui devait permettre d'identifier à quelle série appliquer quel modèle de prévision intermittent. La classification impliquait 4 types de séries intermittentes : les séries lisses, intermittentes, erratiques et grumeleuses.

Toutefois, cette classification fut critiquée par Kostenko and Hyndman (2006) qui ont argumenté qu'il n'y avait en fait que deux classes de séries : celles que la méthode de Croston ajuste mieux et celles que la méthode SBA ajuste mieux. La classification est représentée sur la figure 2.1.

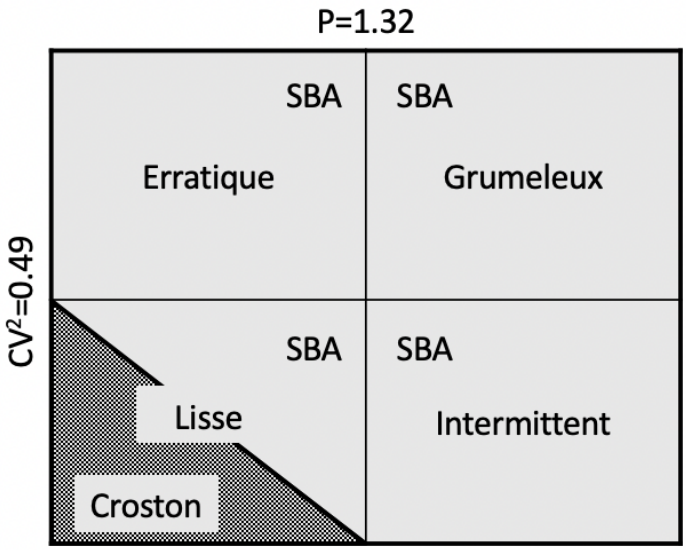

Figure 2.1 Classification des séries d'après la classification de (Syntetos et al., 2005) et de (Kostenko and Hyndman, 2006)

Où p représente le nombre de périodes nulles consécutives moyen dans la série et $CV^2 = \frac{\sigma^2}{\mu^2}$, le coefficient de variation avec σ l'écart-type de la demande et μ, la demande moyenne. Toutefois, cette classification théorique ne semble pas en mesure, en pratique, d'identifier quel modèle performera le mieux sur une série (Kourentzes, 2014), (do Rego and de Mesquita, 2015). Une approche expérimentale est donc à privilégier.

D'un point de vue pratique, il est reconnu depuis Gardner (1990) qu'il est important d'évaluer l'utilité des prévisions en les évaluant selon le contexte dans lequel elles sont utilisées. Or, évaluer les modèles de prévision dans leur contexte d'utilisation requiert un travail supplémentaire long et fastidieux pour bâtir un simulateur capable de recréer ce contexte. Ainsi, établir le lien qui existe entre les métriques d'erreur et les performances de gestion d'inventaire est d'une grande importance. Quelques études ont établi une corrélation positive entre

les performances en gestion d'inventaire et les métriques d'erreur (Syntetos et al., 2010a) (Sanders and Graman, 2009). Cependant, ces études ont été effectuées avec des séries à profil de demande lisse. Dans le cas intermittent, la relation est plus complexe puisque les performances en gestion d'inventaire ne s'accordent pas avec les résultats sur la précision (Solis, 2015), (Kourentzes, 2013), (Teunter and Duncan, 2009). Il n'existe pas pour l'instant d'explication qui permet d'expliquer ce résultat. Les résultats additionnels apportés par cette thèse proposent une explication pour expliquer de tels résultats au chapitre 7.

Un aspect peu étudié de la sélection de modèles est le choix de l'agrégation des résultats. Par exemple, un modèle de prévision unique aux performances globales supérieures peut être sélectionné. Autrement, chaque série peut faire l'objet d'une sélection individuelle de modèles de prévision résultant en la sélection multiple de modèles de prévision. L'impact de la sélection globale versus multiple de modèles de prévision obtient des résultats mitigés selon le profil de demande des séries. Par exemple Tashman and Kruk (1996) et Hyndman et al. (2002) ont montré une meilleure précision pour la sélection multiple sur un ensemble de séries à profil lisse, tandis que Kourentzes (2014) a conclu l'inverse dans un contexte de gestion d'inventaire pour des séries intermittentes. Ces résultats révèlent le besoin d'information additionnelle sur les deux aspects énoncés : 1. la relation entre les métriques de sélection et les performances de gestion d'inventaire et 2. l'impact de l'agregation en contexte de gestion d'inventaire. Ces questions seront approchées dans cette thèse, au chapitre 5.

2.4 Conclusion

Les domaines de la gestion de l'inventaire et des prévisions de la demande ont traditionnellement été étudiés séparément (Prak and Teunter, 2019). Alors que les connaissances s'accumulent, de plus en plus de chercheurs constatent que ces problèmes doivent être approchés conjointement afin d'améliorer les performances d'inventaire de manière significative. L'état de l'art a démontré que plusieurs approches de prévision de la demande et de contrôle des stocks existent et sont toujours développées à ce jour. Pourtant, Syntetos et al. (2016) affirment que malgré plusieurs développements théoriques importants, très peu se sont traduits par des solutions opérationnelles ou ont été intégrés dans des systèmes d'aide à la décision. Ils affirment également que plusieurs heuristiques simples et robustes, éprouvées par l'expérience, mais sans fondements scientifiques ou théoriques, gagnent en popularité dans les applications industrielles. Ces solutions sont difficiles à battre en pratique. Cette information combinée au constat que les gestionnaires ajustent et modifient souvent les prévisions de la demande et arrivent ainsi à améliorer les performances de gestion d'inventaire (Van Donselaar et al., 2010), (Syntetos et al., 2010a) démontre un besoin d'information additionnelle

pour comprendre les facteurs d'influence des performances de la gestion de l'inventaire.

Comme piste de solution à ce problème, plusieurs chercheurs commencent à explorer les liens entre les performances d'inventaire et les prévisions de la demande. Celles-ci étant souvent à la base du processus de prise de décision (Prak and Teunter, 2019). Par exemple, Strijbosch et al. (2011) ont étudié sur des données simulées l'impact de prévisions optimales sur la gestion du stock. Tiacci and Saetta (2009) ont mesuré par simulation les performances de différents modèles de prévision pour différentes conditions de réapprovisionnement avec contraintes de remplissage sur les livraisons et ont conclu que de chercher une métrique pour approximer les performances d'inventaire n'était pas une solution envisageable étant donnée la quantité de facteurs importants ayant une influence sur l'approvisionnement. Plus récemment, Kourentzes et al. (2019) ont optimisé un modèle de lissage SES de manière à minimiser les coûts d'approvisionnement et Bruzda (2020) a optimisé des prévisions avec des contraintes de niveau de service sous plusieurs conditions et hypothèses.

Dans cette thèse nous tenterons également d'identifier quelques facteurs ayant un impact sur les performances d'inventaire. En particulier, trois facteurs ont été identifiés dans la revue de littérature comme nécessitant de la recherche additionnelle : les critères permettant d'évaluer les métriques de performance (chapitre 4), l'agrégation des performances dans la sélection de modèles (chapitre 5) et l'impact des politiques dynamiques versus statiques (chapitre 6). Dans un contexte de recherche où la relation entre les prévisions de la demande et la gestion de l'inventaire est de plus en plus au coeur des activités de recherche du domaine. Ces facteurs à l'étude dans la thèse auront pour but d'aider la conception d'un système de gestion de l'inventaire pour un portefeuille de produits à profil de demande mixte.

La prochaine section discute de notre approche pour faire l'étude de ces facteurs.

CHAPITRE 3 DÉMARCHE ET ORGANISATION

3.1 Introduction

Les chapitres précédents ont introduit le contexte de recherche, la problématique ainsi que les outils et limites des connaissances actuelles sur le sujet de recherche.

Le constat est que le projet de recherche s'inscrit dans un contexte où l'accessibilité des technologies issues de l'intelligence artificielle est de plus en plus facile et où le développement continu et rapide de ces technologies permet aux techniques de prévision et de prise de décision automatisée de s'améliorer continuellement. Un système de gestion de l'inventaire développé dans ce contexte devrait donc être modulaire pour pouvoir s'adapter et intégrer rapidement sans trop d'efforts ces nouvelles technologies. C'est pourquoi le développement méthodologique d'évaluation et de sélection des techniques de prévision et décision est un problème de premier plan dans la recherche sur les systèmes de gestion de l'inventaire. Spécifiquement dans les cas où des demandes de différents profils sont présentes dans les données, car différents profils requièrent différents traitements. Certains points en particulier requièrent de la recherche additionnelle. Ces points sont formulés sous la forme d'objectif et de questions de recherche à la section 3.2. Cette section sera suivie par la présentation du contexte industriel (section 3.3) et de la méthodologie (section 3.4).

3.2 Les objectifs de recherche

Étant donné le contexte présenté en introduction, l'objectif général du projet de recherche est de concevoir un système de gestion d'inventaire pour un portefeuille de produits à profil de demande mixte.

Une approche pour atteindre l'objectif serait de développer un système autonome composé de 8 opérations séquentielles. Le système et ses composants sont présentés sur la figure 3.1. Ils consistent essentiellement en deux ensembles d'opérations dont les variables d'entrées à la période t sont la demande (Y_t) ainsi que d'autres variables exogènes, de suivi d'inventaire et métadonnées utiles (X_t). La sortie est la variable de décision (u_t).

Le premier ensemble d'étapes sert à obtenir des prévisions de la demande fiables. Il s'agit pour ce faire de générer des prévisions de la demande avec plusieurs modèles de prévisions (opération 1), d'évaluer leur erreur d'ajustement (opération 2), d'assurer par un contrôle statistique la qualité des prévisions et de prendre des mesures correctives en cas de déviation

significative (opération 3). Une fois les prévisions prêtes, les meilleurs modèles sont sélectionnés (opération 4). L'accomplissement de l'ensemble de ces tâches garantira des prévisions de la demande fiables.

Les prévisions et les autres données serviront ensuite à alimenter le deuxième ensemble d'étapes dont l'objectif est la prise de décision de réapprovisionnement. Des politiques d'inventaire prendront donc les décisions de réapprovisionnement (opération 5). Les performances sur les données disponibles seront évaluées (opération 6) et un contrôle des performances sera effectué pour assurer un ajustement correct des politiques (opération 7). Finalement, les politiques les plus performantes seront sélectionnées pour la prise de décision (opération 8).

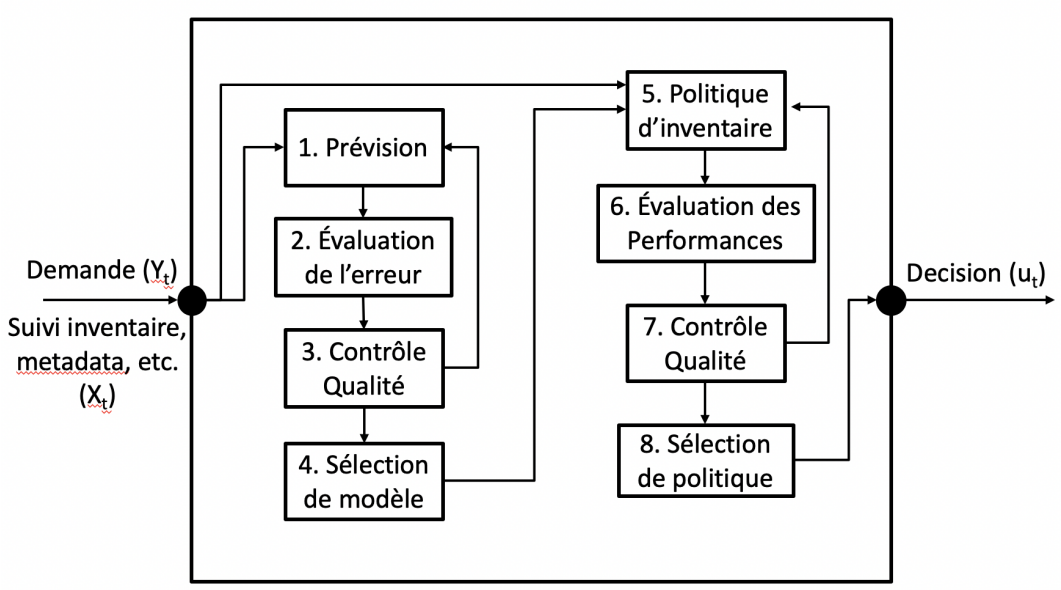

Figure 3.1 Méthodologie

Des contributions scientifiques liées au développement de certaines composantes de ce système sont apportées dans cette thèse. Notamment au niveau du développement méthodologique sur le point 2, où l'absence de méthodologie d'évaluation des métriques de performances basées sur l'erreur est comblée. Il en sera question au chapitre 4.

Suivant la revue de littérature, il apparaît que l'impact de l'agrégation dans la sélection de modèles n'est pas bien compris. De plus, le lien entre la précision dans les prévisions et leur utilité en gestion d'inventaire n'est pas clair. Certains auteurs suggèrent une relation positive, alors que d'autres ne voient pas de relation. Le chapitre 5 apporte une contribution méthodologique sur la sélection de modèles multiples en contexte de gestion d'inventaire (point 4). Ainsi, à partir des métriques les plus sensibles et fiables sur notre ensemble de données, on étudiera l'impact de l'agrégation dans la sélection de modèles sur les performances d'inventaire.

Les résultats des expérimentations apportent également des connaissances additionnelles sur la relation entre la précision des prévisions de la demande et les performances en termes de gestion d'inventaire en présence de séries à profil mixte.

Finalement, le chapitre 6 porte sur la prise de décision et apporte une contribution méthodologique en proposant une approche basée sur le cadre conceptuel de l'apprentissage par imitation pour optimiser une politique d'inventaire dynamique adaptative au changement de performance des modèles de prévision de la demande (point 5).

Afin de pouvoir répondre à ces questions, d'autres tâches ont été accomplies notamment les points 1 et 6. Leurs détails ainsi que notre approche de résolution sont présentés à la section 3.4. Une des motivations de ce projet de recherche en plus de sa motivation scientifique provient du milieu industriel. Le contexte industriel qui constitue également le cas d'étude dans chacune des contributions est présenté dans la section suivante.

3.3 Contexte industriel

Les données utilisées proviennent de notre partenaire industriel Logistik Unicorp (LU), qui fournit un service de gestion du programme d'uniforme pour plusieurs organisations à travers le pays et ailleurs dans le monde. LU fournit le service de la conception à la distribution pour chaque membre individuel des organisations clientes.

LU doit conserver en inventaire les items des uniformes pour être en mesure de répondre à 95% de la demande en tout temps. La demande provient directement des membres individuels des organisations clientes qui passent leur commande via une application web. La demande réelle est donc accessible pour prévoir et organiser les opérations.

Les commandes sont passées individuellement pour chaque item et plusieurs commandes ouvertes en même temps sont acceptées par les fournisseurs. L'entreprise possède un système de suivi de l'inventaire continu avec un MRP pour suivre et faire des recommandations d'achat aux acheteurs. Les prévisions et décisions de réapprovisionnement sont agrégées en semaine et, selon le fournisseur, les commandes doivent couvrir la demande durant R périodes (semaines). Les fournisseurs sont engagés à livrer les commandes à l'intérieur d'une période de LT semaines considérée comme étant le délai de livraison. Les livraisons aux clients individuels sont gérées par un fournisseur externe. La structure de la chaîne logistique dans laquelle s'insère le cas d'étude est présentée sur la figure 3.2.

Les données qui ont été fournies sont la demande par unité de gestion des stocks (SKU) quotidienne issue des requêtes de chaque client depuis l'application web. Les informations sur chaque SKU comprenaient le LT, R, ainsi que le groupe de produit, la saison et le sexe

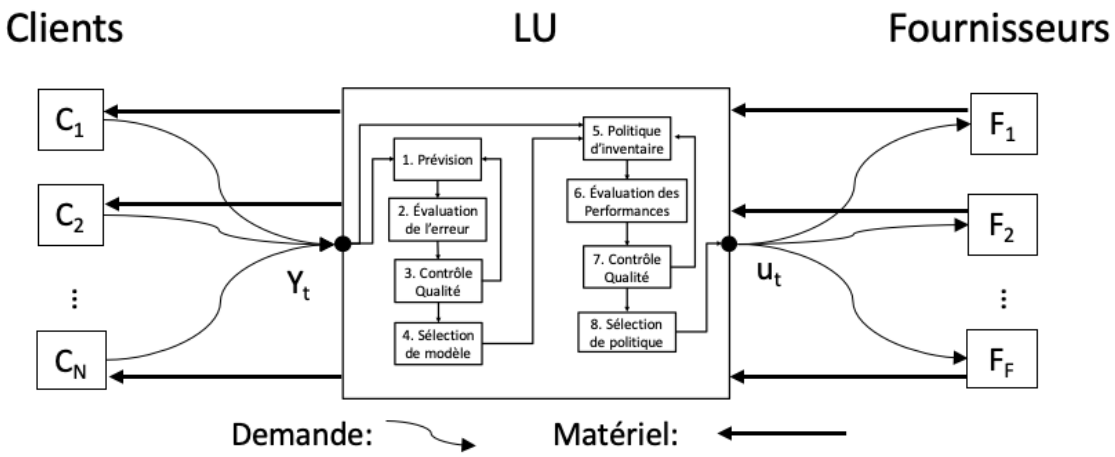

Figure 3.2 Structure de la chaîne Logistique

du produit. On disposait de la demande sur tous les items depuis 2012 jusqu'au début 2019. Pour plus de 6 millions de commandes passées par un peu plus de 1,5 million de clients sur un peu moins de 10k SKU. En utilisant la classification théorique basée sur les caractéristiques des séries de (Syntetos et al., 2005), on obtient la distribution des profils des séries présentée à la figure 3.3.

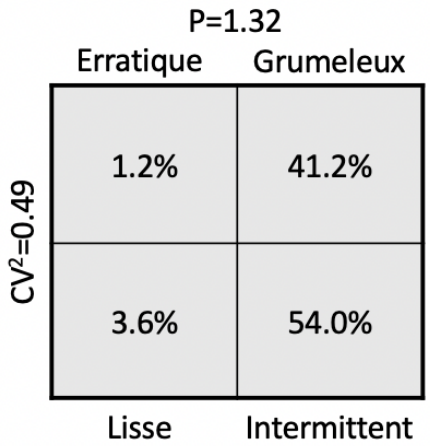

Figure 3.3 Classification des SKU d'après la classification de (Syntetos et al., 2005)

La section suivante présente notre méthodologie de résolution du problème dans le contexte de recherche présenté avec pour objectif industriel d'aider l'entreprise à améliorer la gestion de son inventaire.

3.4 Méthodologie

La méthodologie pour accomplir les trois objectifs de recherche et l'objectif principal a requis le développement de deux systèmes sur lesquels les expérimentations pour les contributions scientifiques reposent. Le premier système est un système pour générer des prévisions de la demande et le second est un simulateur de la gestion de l'inventaire. La figure 3.4 représente les deux systèmes avec leurs interactions. Le détail des opérations effectuées par les deux systèmes est présenté dans les sous-sections suivantes.

3.4.1 Système de prévision de la demande

Le système de prévision et d'évaluation des prévisions se trouve à gauche de la figure 3.4. On y représente les paramètres entrants et les structures de données résultantes.

Étape 1 - Préparation des données : À partir de la demande quotidienne, on sélectionne une période d'agrégation (agg). Pour l'ensemble des expérimentations, ce paramètre est gardé constant pour une agrégation hebdomadaire étant donné qu'il s'agit de l'agrégation utilisée par le système du partenaire. La somme de la demande de tous les clients pour un item sur une période est agrégée afin de transformer la demande en séries temporelles de périodes agg.

Étape 2 - Partitionnement : Avec les données agrégées au bon niveau de granularité, on partitionne les données en trois groupes. Pour ce faire, deux paramètres sont requis : t et t_v. Le paramètre t donne la période avant laquelle les données sont utilisées pour l'entraînement des modèles de prévision. Les périodes supérieures à t sont utilisées comme données test. On fait évoluer t de façon dynamique jusqu'à la période t_v. Cette période indique la première période de validation. C'est-à-dire que la demande à partir de t_v n'est utilisée que pour mesurer les performances des modèles.

Étape 3 - Prévision : Les données d'entraînement (précédent t) sont utilisées pour paramétrer les modèles de prévision sélectionnés. Notre système permet de prendre n'importe quelle fonction de prévision à laquelle on applique chaque série temporelle dans nos expérimentations (arima avec composante saisonnière, SBA, TSB, snaive et theta) ont été utilisés, car ils représentent bien les divers types de modèles de prévision statistique utilisés.

Étape 4 - Évaluation : Les prévisions peuvent être évaluées avec des métriques de performances basées sur l'erreur. Pour ce faire, les données tests (entre t et t_v) sont utilisées.

À partir de cette étape, le système permettant de générer et d'évaluer des prévisions de la demande est complet. Ce système servira de cadre de base pour nous permettre de répondre à la première problématique de recherche. Les deux autres problématiques requièrent, en plus

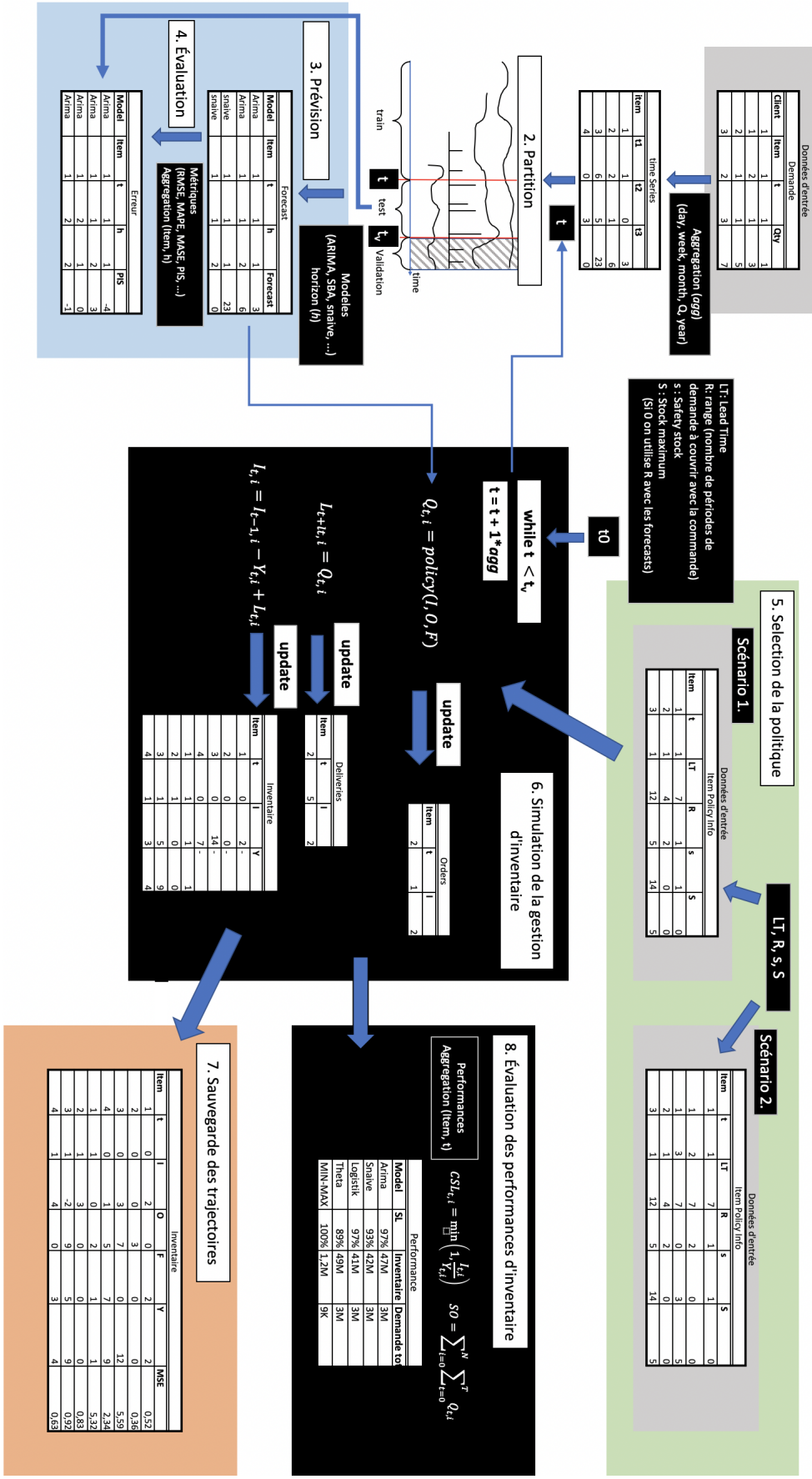

Figure 3.4 Entrants, extrants et paramètres des systèmes implémentés

du système de prévision, un simulateur du processus de réapprovisionnement de matériel. Ce système permet d'évaluer les performances en termes de gestion d'inventaire. Les étapes suivantes en présentent le détail.

3.4.2 Simulateur de la gestion d'inventaire

Étape 5 - Sélection de la politique : Une fois les données préparées, on sélectionne les paramètres de la police qu'on souhaite utiliser. Les paramètres à sélectionner sont : le délai de livraison (LT), la période de couverture (R), le point de recommande (s) et le niveau de réapprovisionnement (S) pour chaque item. Les paramètres peuvent être dynamiques et modifiés chaque période. Ainsi, toutes les politiques à suivi continu peuvent être générées. Dans le cas où S est paramétré à une valeur de 0, alors les prévisions de la demande issues du système de prévisions sont utilisées pour prendre la décision de réapprovisionnement. Divers scénarios avec diverses configurations de politiques dynamiques ou statiques peuvent ainsi être générés.

Étape 6 - Simulation de la gestion de l'inventaire : La simulation démarre en initialisant des tables contenant pour chaque item l'inventaire initial I, les commandes en cours O et les livraisons L. Dépendamment de la politique, les prévisions de la demande à la période t sont générées avec les modèles choisis pour chaque item sur l'horizon requis. Utilisant la politique spécifiée à la période t une décision de réapprovisionnement Q est prise. La demande actuelle Y et les livraisons à la période t sont ensuite additionnées à l'inventaire. L'étape est répétée pour un nombre spécifié de périodes ou jusqu'à atteindre la période t_v.

Étape 7 - Sauvegarde des trajectoires : Les états de chaque itération des simulations comprenant les informations sur les quantités en stock I, les quantités en cours de commande O, les prévisions de la demande F, la demande Y, et l'erreur d'ajustement (sur l'échantillon d'entraînement) $MSE_{in-sample}$, sont sauvegardées dans une table. L'objectif de sauvegarder les états observés à chaque itération de la simulation est de pouvoir répondre à la troisième problématique de recherche.

Étape 8 - Évaluation des performances d'inventaire : Les résultats des simulations sont analysés une fois les itérations terminées. Le niveau de service et le stock total commandé sont les métriques utilisées. Plus de détails sur ces métriques sont donnés dans les chapitres 5 et 6.

3.5 Conclusion

Ce chapitre a présenté en détail les objectifs de recherche, le contexte industriel dans lequel s'inscrit le cas d'étude de la thèse ainsi que la méthodologie utilisée pour accomplir les objectifs. Pour la suite, les chapitres 5 et 6 utilisent les deux systèmes présentés comme cadre de base pour conduire les expériences. Les 3 chapitres suivants expliquent avec plus de détails les méthodologies spécifiques à la résolution des problématiques de recherche.

CHAPITRE 4 ARTICLE 1: A METHODOLOGY TO EVALUATE FORECASTS PERFORMANCE METRICS

St-Aubin, P., Agard, B.

Abstract *- Thus far, work on performance metrics has been done without knowing the ground truth on the real performance of forecasting models. This paper proposes a new methodology to measure the sensitivity and reliability of forecasts performance metrics. The methodology is tested using multiple time series of different scales and demand patterns, such as intermittent demand. The idea is to add to each series a noise following a known distribution to represent forecasting models of a known error distribution. Varying the parameters of the distribution of the noise allows to evaluate how sensitive and reliable performance metrics are to changes in bias and variance of the error of a forecasting model. The experiments concluded that sRMSE is more reliable than MASE in most cases on those series. sRMSE is especially reliable for detecting changes in the variance of a model and sPIS is the most sensitive metric to the bias of a model. sAPIS is sensible to both variance and bias but is less reliable.*

Keywords: Performance Measurement, Forecasting, Intermittent demand, Forecast Accuracy, Time Series

4.1 Introduction

The forecasting community has long been searching for the best method to assess the performance of forecasting models. This search was in part driven to solve the difficulties of forecasting multiple series of different scale and demand patterns such as intermittent demand (Croston, 1972). Makridakis et al. (2018) showed that the best forecasting techniques at the last M competition presented a small difference in performance. Therefore, correctly identifying the best technique is going to become increasingly difficult as techniques get closer to perfection. For this reason, evaluating with certitude what the sensitivity and reliability of performance metrics is going to become an important factor for the selection and ranking of forecasting models.

This paper addresses this problem by proposing a new methodology to measure the sensitivity and reliability of performance metrics. It presents how variation in bias and variance of the error of a forecasting model influences the performance obtained with a specific metric. This is done by comparing performance across multiple time series of different scale and demand pattern.

The proposed methodology allows one to identify on a given dataset what is the most sensitive and reliable performance metric. This is important since the community is aware of the difficulties of selecting appropriate parameters and the best models, especially in the context of intermittent demand (Kourentzes, 2014).

The sections in this paper are divided as follows. Section 4.2 presents previous work concerning performance metrics and evaluation of performance metrics. Section 4.3 presents the methodology to measure sensitivity and reliability. Section 4 presents the results of the application of the methodology on real data. Finally, section 5 provides a summary and recommendations based on the empirical results.

4.2 Previous Work

In many industrial applications such as retail, forecasting performance must be aggregated to avoid the high complexity of analyzing performance for every single product (Hoover et al., 2009). In this case, it is important to assess the performance of models in the best possible way, since it will translate into actual and/or opportunity losses (Makridakis and Hibon, 2000).

In some cases, demand comes in an intermittent or erratic fashion (Croston, 1972). This type of demand presents a high number of consecutive null demands, or it can be characterized with a high coefficient of variation (Syntetos and Boylan, 2005). These cases present a challenge in measuring the performance of forecasting models due to the high number of null demands (Solis, 2015), which can cause some performance metrics to be undefined.

Let us note Y the time series with Y_t the data used to estimate the model (in-sample), Y_h the hold out sample (out-of-sample) data to estimate performance and F_h the forecast associated with Y_h. The error is noted $e_h = Y_h - F_h$ the error value at horizon h. Different operations can be performed on the error e_h to assess the forecasting performance with greater precision and discernment.

Armstrong and Collopy (1992) explored those operations through empirical comparisons. They were not able to measure sensitivity in absence of ground truth but mentioned the importance of the sensitivity of performance metrics. they measured reliability as being the average Spearman correlation for pairwise comparisons among five different subsample. They recommended Median Absolute Percentage Error (MdAPE) to select the most accurate forecasting method and they also introduced relative based error metrics such as Relative Absolute Error (RAE). This metric scales the absolute error with the one of a naïve method, but (Syntetos and Boylan, 2005) argued that such a metric along with the Mean Absolute

Percentage Error (MAPE) were not appropriate for intermittent demand since they could involve division by zero.

To solve problems of relative based scaling methods, Hyndman and Koehler (2006) introduced the Mean Absolute Scaled Error (MASE), an adaptation of the relative based metric which scales the error using the in-sample error of a benchmark method instead of the out-of-sample error of a benchmark method.

Wallström and Segerstedt (2010) studied the case of forecast evaluation for intermittent case. They introduced cumulative methods such as Period in Stock (PIS) and the Cumulated Forecast Error (CFE) as they are not biased toward zero forecast compared to other classical metrics such as RMSE and MAD (Teunter and Duncan, 2009).

Then Petropoulos and Kourentzes (2015) used the idea proposed by Hyndman and Koehler (2006) to scale MSE, MAE and PIS with the in-sample mean demand instead of the in-sample error of a benchmark method. A similar idea was used in (Billah et al., 2006) but was critiqued in (Hyndman and Koehler, 2006) as the in-sample mean could be skewed in presence of non-stationary data.

Previous work on performance metrics was first to find metrics able to compare the performance of models across multiple series of different scales. The second motivating factor for research on performance metrics was to ensure definite and stable metrics across all cases that could be met like for the case of intermittent demand.

So far, no research has primarily focus on the sensitivity and reliability of performance metrics to detect which metric should be used to select and optimize a forecasting model. The next section present a new methodology that allows one to quantify the sensitivity and reliability of a performance metric.

4.3 Methodology

This section presents a new methodology allowing one to compare performance metrics according to their sensitivity and reliability to changes of the error distribution of a forecasting model applied to multiple time series. The main idea of the methodology is to build fictitious forecasting models with a chosen error distribution. Allowing one to rank the models exactly given their error distribution is known.

The methodology is divided into four main sections (see Figure 4.1): 1. build forecasting models with known errors, 2. select performance metrics to evaluate, 3. measure sensitivity and 4. measure reliability.

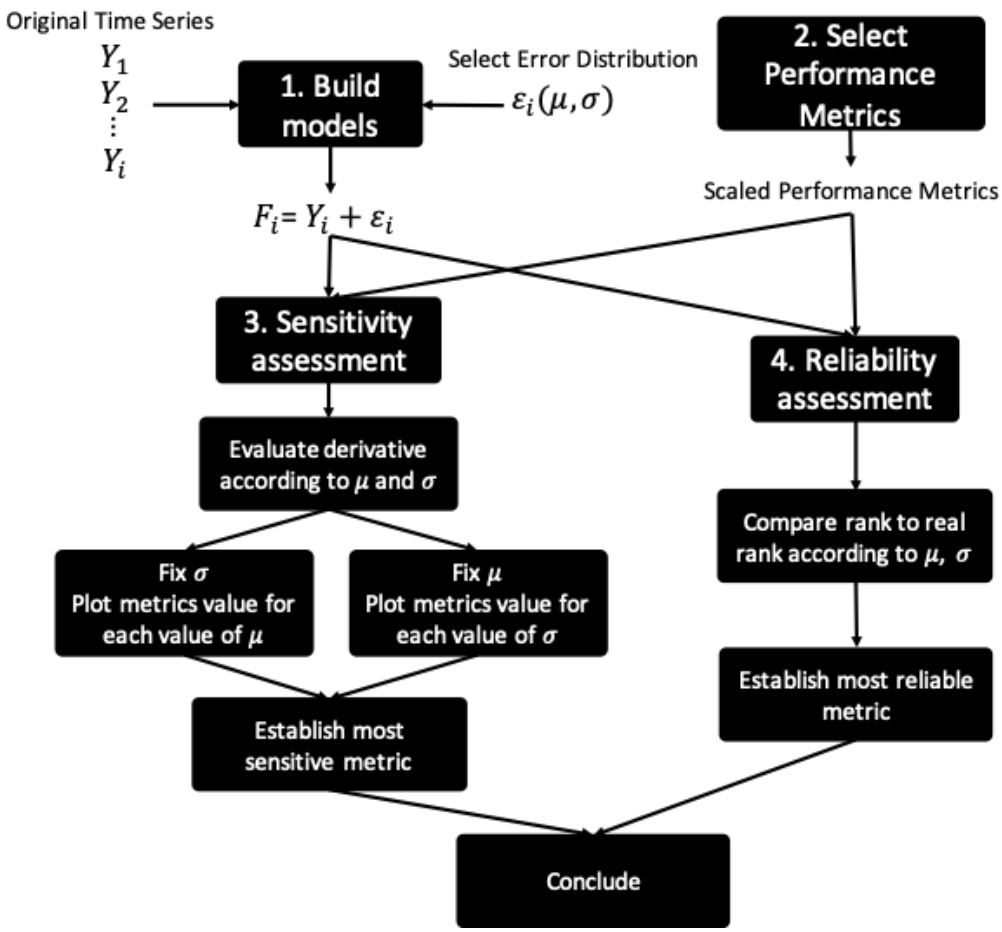

Figure 4.1 Experiment diagram

4.3.1 Build Models

Here forecasting models are defined as the original time series to forecast to which a random noise is added:

$$F_{i,h} = Y_{i,h} + \varepsilon_i \tag{4.1}$$

Where $\varepsilon_i \sim N(\mu_i, \sigma_i^2)$ is the random noise.

The choice of the error distribution could have been any distribution and the following results would still be valid. Indeed, since to measure the performance of a forecasting model on multiple time series, the aggregation of the performance must follow a normal distribution according to the central limit theorem. Thus, to simplify the following analysis, the normal distribution is chosen so that the bias and variance of the error distribution of the forecasting models are directly related to the chosen parameters of the error distribution.

i represents the index of the time series and h the horizon of the forecast.

In a case with multiple time series, the parameters of the distribution of the error of a model should consider the different scales of the different series.

To do so, the chosen parameters of the error are coefficients of the in-sample mean value of each series, making the error size relative to the scale of the series.

$$\mu_i = \alpha \frac{1}{N} \sum_{t}^{N} Y_{i,t} \tag{4.2}$$

$$\sigma_i = \beta \frac{1}{N} \sum_{t}^{N} Y_{i,t} \tag{4.3}$$

Where $\frac{1}{N} \sum_t^N Y_{i,t}$ is the in-sample mean value of the time series. The parameters α and β take values between 0 and 1 so the parameters of the error distribution are equal to a proportion of the in-sample mean.

To summarize, the fictitious forecasting models are composed of the actual value plus a noise that follows a chosen normal distribution. For each fictitious model, the parameters of the error distribution are set to a certain proportion α and β of the in-sample mean of each series. This makes the error distribution proportionate to the scale of each series.

This way, one can create multiple fictitious models varying α and β in an ordered manner with different increments to measure how sensitive and reliable performance metrics are to detect the difference in bias and variance of the error of different models. In further sections, the models are defined by the parameters of their error distribution which are set to a proportion of the in-sample mean.

4.3.2 Select performance Metrics

In this section one simply needs to select the performance metrics he wants to evaluate. For the experiment presented in section 4.4, $RMSE$, $sMAPE$, $MASE$ and PIS were chosen since they are all common and known metrics. In addition, they can be, or are, adapted to always be defined and can be scaled using the (Petropoulos and Kourentzes, 2015) scaling factor. Also, the results by (Wallström and Segerstedt, 2010) seemed to show that all metrics of the same category are strongly correlated, and for this reason, only one metric of each metric class is used.

Scaled versions of the metrics are used: $RMSE$ ($sRMSE$), PIS ($sPIS$) and Absolute PIS ($sAPIS$). They are scaled to allow comparison of performance across series of different scales.

The metrics, along with their scaling factors, are described under:

$$sRMSE = \frac{\sqrt{\frac{1}{H}\sum_h^H e_h^2}}{\frac{1}{N}\sum_t^N Y_t} \tag{4.4}$$

$$sPIS_H = \frac{-\sum_h^H \sum_{i=1}^h e_i}{\frac{1}{N}\sum_t^N Y_t} \tag{4.5}$$

$$sAPIS_H = \frac{|\sum_h^H \sum_{i=1}^h e_i|}{\frac{1}{N}\sum_t^N Y_t} \tag{4.6}$$

We used scaled metrics as defined in (Petropoulos and Kourentzes, 2015) as it is easy to interpret in practical terms. It will also allow us to measure what is the impact on precision and reliability of using a scaling factor robust to non-stationary data as it is for $MASE$. Finally, the aggregation of results across both time series and different horizons is done by taking the mean. For this reason, an additional "m" was added to the metrics abbreviation so that one can make a distinction between the aggregation of horizon periods and the series aggregation.

4.3.3 Sensitivity

Variation of the bias and the standard deviation parameter will reveal how different metrics react to changes of these parameters. This reveals which metric is the most sensitive and potentially the best one to distinguish between two models of similar error distributions.

The same thing can be done by fixing one parameter for several different values and varying the other one. That will allow one to estimate the influence of each parameter on the other one.

The results will allow one to draw conclusions about the sensitivity of metrics to standard deviation and bias of a model. The second measure of sensitivity presents how stable sensitivity is to one parameter given changing values for the other one.

4.3.4 Reliability

In this section reliability is evaluated in terms of ranking of models. Since the real error distribution of the models are known, one can rank the models exactly. So, the ranks obtained with the performance metrics can be compared to the real rank of each model.

To test this, different configurations of scales for the bias, the standard deviation and the

scale of the variation ($\Delta\mu$ and $\Delta\sigma$) between two models are considered.

For example if F^1 has the parameters (μ, σ) the closest model in terms of error would be F^2 of parameters $(\mu, \sigma + \Delta\sigma)$.

The different situations are presented in 4.1. The plus and minus symbols represent the relative order of magnitude of the different parameters in comparison to the others. For example, if μ and σ both have a minus symbol, α and β would be less than 0.01. Which means μ and σ would be less than the order of magnitude of 1% of the in-sample mean. Otherwise, this would mean they would both be of the order of magnitude of 1% of the in-sample mean. The same applies for $\Delta\sigma$ and $\Delta\mu$, which represent the difference in standard deviation and bias of the error distribution of the forecasting models. A minus symbol would indicate α and β are two orders of magnitude less than 1% and the plus symbol indicates they are an order of magnitude less than 1%. The Δ must be smaller than the parameter so that, once multiplied by the number of models, it reaches the same order of magnitude as the parameter. For example, the (+++0) configuration implies σ, μ of the order of 1% and $\Delta\sigma$ of 0.1%. So that $\Delta\sigma$ multiplied by the number of models is of the order of 1%. The (++-0) configuration would have both parameters of the order of 1% and $\Delta\sigma$ of 0.01%.

Table 4.1 Configurations of standard deviation, bias and variation standard deviation and bias to test and evaluate the reliability of metrics

σ	μ	$\Delta\sigma$	$\Delta\mu$
-	-	-	0
+	-	-	0
+	-	+	0
-	+	-	0
+	+	-	0
+	+	+	0

To rank the models exactly given their error distribution, only one of the distribution parameters at a time will vary. So, either the bias of the standard deviation will be fixed, while the other will change. Table 4.1 presents half the configurations, where the others will consider a variation of bias with fixed standard deviation. Note there is no configuration where σ is small and $\Delta\sigma$ is large. This is because large variations would bring the interval of σ from small to large, making it the same configuration as both $+\sigma$ and $+\Delta\sigma$.

The idea is to measure, for each configuration of bias, standard deviation and delta, the quality of the ranking rendered by the performance metrics. The quality is obtained by comparing metrics rankings to the real ranking of models.

Two different metrics are used to measure the quality of the rankings. The first one is

Spearman's rank correlation (Spearman, 1987). This metric returns the correlation between the real rank and the rank given by metrics. The measure's value lies between -1 and 1 to indicate negative to positive perfect correlation. It allows us to make conclusions about the general tendency of metrics to rank models in the same order as real rankings.

The second one is the normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002). This metric provides a score between 0 and 1, 1 being the perfect ranking. The main difference between nDCG and Spearman's rank correlation is that nDCG gives more importance to ranking models correctly in the firsts rank compared to those in the lasts rank.

Indeed, the nDCG metric provides a relevance score to every model it needs to rank and it discounts the relevance, as it appears lower in the ranking. In this experiment, 1 over the real rank of each model is used as their relevance score. This is especially important given that being right for the ranking of the first N models is of greater importance than showing a similar general tendency, as does Spearman's rank correlation.

Reliability is observed in function of the number of forecast horizons and the number of series performances used to average performance as illustrated in Figure 4.2. The figure presents three different series. Its global performance is calculated by varying the number of forecast horizons used to estimate the performance for a series (red square), and by varying the number of series performances that are averaged (blue square).

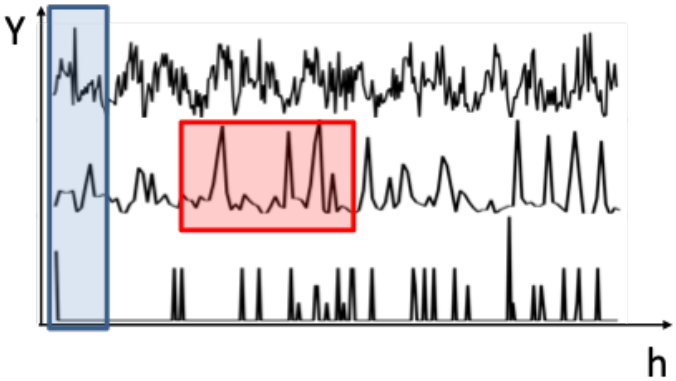

Figure 4.2 Representation of the averaging variables to estimate performance

To make sure the results were not obtained only by chance, this step is repeated 35 times with different series and the results are averaged. The average reliability of all repetitions is finally plotted to visualize the behaviour of the metrics. Thus, the convergence and the general superiority in terms of correlation and nDCG is observed to conclude on the relative reliability of the metrics. So even if some observations of the correlation are not formally significant in a statistical sense, we can still conclude on the overall performance of the metric

compared to the others.

4.4 Results

4.4.1 Data

The data used to test the methodology comes from Logistik Unicorp, a company in charge of supplying uniforms to the members of organizations such as the Canadian Army, Canadian Borders, and many other organizations. Several thousand time series were available. Each one represents the demand for an article of clothing in the uniform of a certain size. The demand was aggregated to weekly demand to fit with the MRP system and for other planning purposes. Series with at least two years of demand were kept. That left 23 weeks of horizon periods to evaluate performance on.

4.4.2 Standard deviation sensitivity

To measure the metric's sensitivity to the standard deviation of a model, the bias of the model was set to $\mu = 0$. This leaves only the standard deviation parameter σ that was studied through the standard deviation of the distribution, which is equal to σ. The parameter varied between 1% and 50% of the in-sample mean.

Figure 4.3 illustrates the global performance of all the metrics values. The dashed line represents the real standard deviation. $msRMSE$ is close to the real standard deviation of the model, which is what is expected by theory. This reassures that the scaling method worked, since the observed trend for $msRMSE$ is almost perfectly aligned with the real value.

All metrics show linear growth. Therefore, the slope of each metric can be approximated with the equation below.

$$vr = mean\left(\frac{L(\sigma + \Delta) - L(\sigma)}{\Delta}\right) \tag{4.7}$$

Where vr is the variation rate, $L(\sigma)$ is the performance metric expressed in function of standard deviation L and Δ is the variation in standard deviation, which is 5% in this case.

This allows conclusions to be drawn on each metric's sensitivity to standard deviation in absence of bias in forecasting models:

Table 4.2 represents the variation rate of metrics in Figure 4.3.

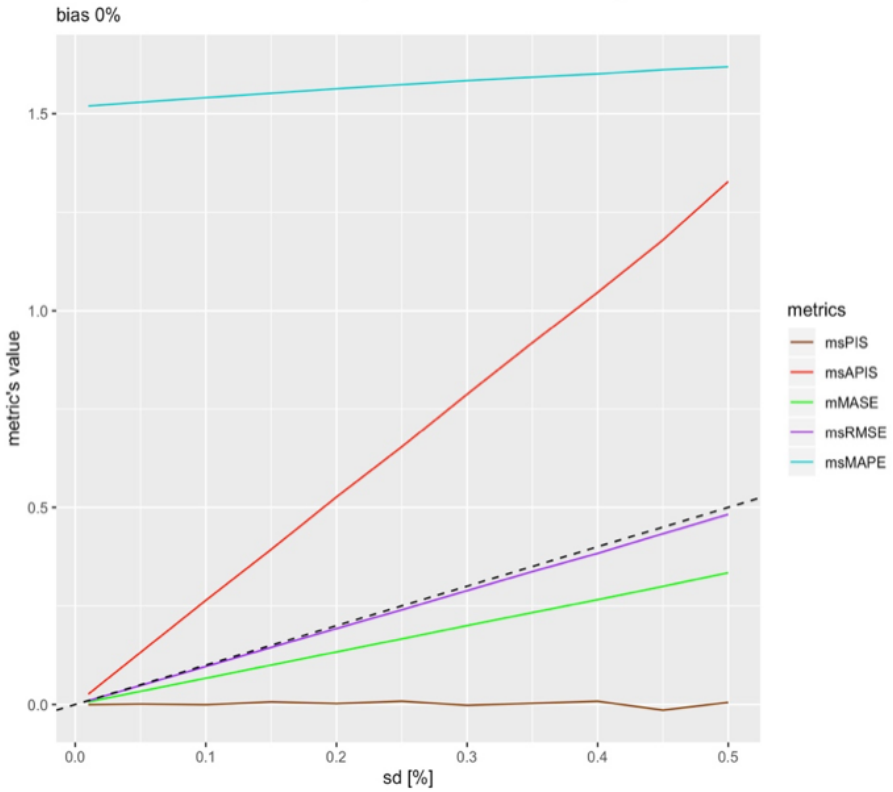

Figure 4.3 Standard deviation sensitivity of scaled performance metrics in absence of bias

Table 4.2 Average variation rate of performance metrics for a variation in standard deviation and fixed bias of forecast models

Metric	msPIS	msAPIS	mMASE	msRMSE	msMAPE
variation rate (vr)	0.01	**2.65**	0.67	0.96	0.21

What is important to note is the order of magnitude of the variation more than the values themselves. Indeed, since the values probably vary from one dataset to another and depending on conditions such as the number of series, the number of forecast horizons, the level of intermittence of the data, etc. $msRMSE$ is the closest to the real standard deviation. $msAPIS$ is the most sensitive metric to change in standard deviation, with a variation rate of an order of magnitude higher than the other. The average of $msPIS$ is near zero, since it has a negative value when standard deviation makes the forecast lower than the actual value. For this reason, it is expected that $msPIS$ is a good estimator of forecast bias as it is not affected by standard deviation. With this first result, $msMAPE$ and $mMASE$ are the first and second least sensitive metrics to variation in the standard deviation of forecasting models.

4.4.3 Bias sensitivity

Let us evaluate bias sensitivity by fixing the standard deviation of forecasting models to 1% and then varying the bias from −50% to +50% of the in-sample mean.

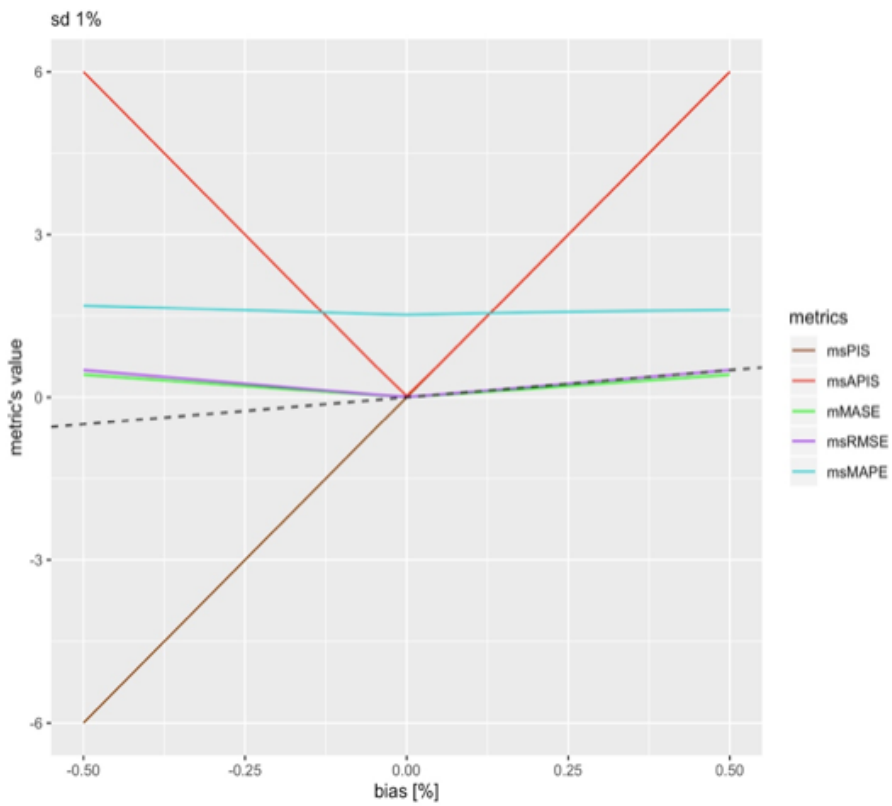

Figure 4.4 Bias sensitivity of scaled performance metrics with 1% standard deviation in forecast models

Figure 4.4 presents the variation of performance metrics for changes in bias of forecasts. Both cumulative metrics seem to be more sensitive to bias. $msRMSE$ and $mMASE$ are close to the real bias values as shown by the dashed line. The variation in the values of metrics is also partly linear. The slope of this linear trend will therefore be estimated in the same way as standard deviation sensitivity, but with an absolute value to remove the sign from the slope:

$$vr = mean\left(\frac{L(\mu + \Delta) - L(\mu)}{\Delta}\right) \tag{4.8}$$

Where $L(\mu)$ is the performance metric expressed in function of bias μ and Δ, the variation in bias, which is 5%.

Table 4.3 represents the average variation rate considering a symmetric rate. It brings a slight

Table 4.3 Average variation rate of performance metrics for a variation in bias and fixed standard deviation of forecast models

Metric	msPIS	msAPIS	mMASE	msRMSE	msMAPE
variation rate (vr)	**12.00**	11.87	0.80	0.95	0.28

simplification to the result in the case of $msAPIS$ and $msMAPE$. The difference in the calculated rate is noticeable when comparing $msAPIS$ with $msPIS$, which does not have this effect around zero since the effect of standard deviation on msPIS is null, on average. (Hyndman and Koehler, 2006) and (Goodwin and Lawton, 1999) noted that $msMAPE$ provides higher penalties to a negative bias than for a positive bias. The difference in rates between the negative and positive bias for $msMAPE$ is around 0.11. Meaning that the variation rate is greater by 0.11 when bias is negative, versus when it is positive.

All metrics have the same order of magnitude in vr except for cumulative metrics, which are of two orders of magnitude higher. $msRMSE$ is the closest metric to real bias.

4.4.4 Standard deviation-Bias sensitivity

4.4.4.1 Standard deviation in function of bias

For the first measurements of sensitivity, bias and then standard deviation, were fixed to a minimal value. This contrasts with situations in real life, in which selection of models probably implies identifying models with different values of bias and standard deviation. This section will therefore study the vr of performance metrics for a change of both standard deviation and bias.

Let us first examine what happens to the vr in the standard deviation direction when changing the bias of the models on Figure 4.5.

Figure 4.5 shows the metric's value in function of a model's standard deviation for different values of bias. The growth in all is linear, except for $msAPIS$, which has two modes. The first mode makes $msAPIS$ behave like $msPIS$. It is the case until standard deviation reaches a high enough value so that an increase in standard deviation has an impact on the metrics value. All other metrics also present two modes, but with the flat mode being much shorter. The difference between the length of the first mode is explained by the variation rate of bias, which is greater for $msAPIS$ than for the other metrics, which have similar variation rates for both standard deviation and bias. $msPIS$ converges to values close to bias and it is not affected by standard deviation. To represent the impact of bias on the standard deviation vr, its variation is plotted for every bias value. To do so, the average vr in standard deviation

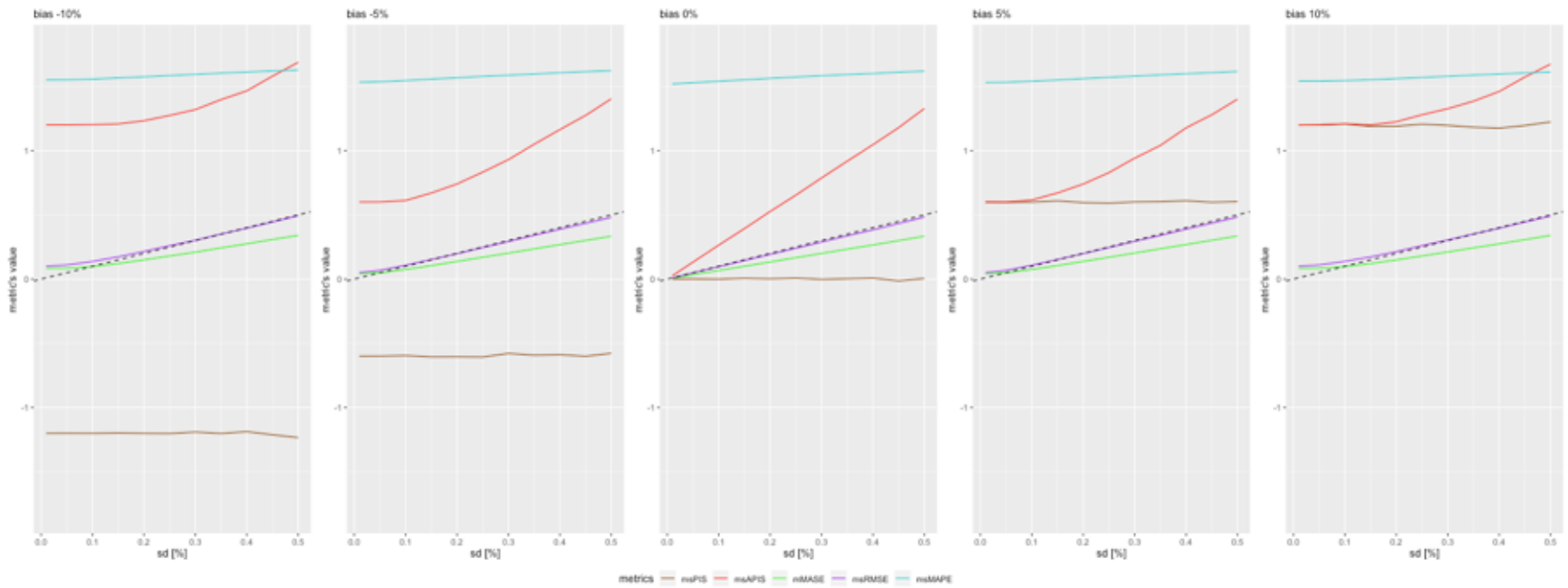

Figure 4.5 impact of change in bias on the different performance metrics

direction of the different metrics is taken.

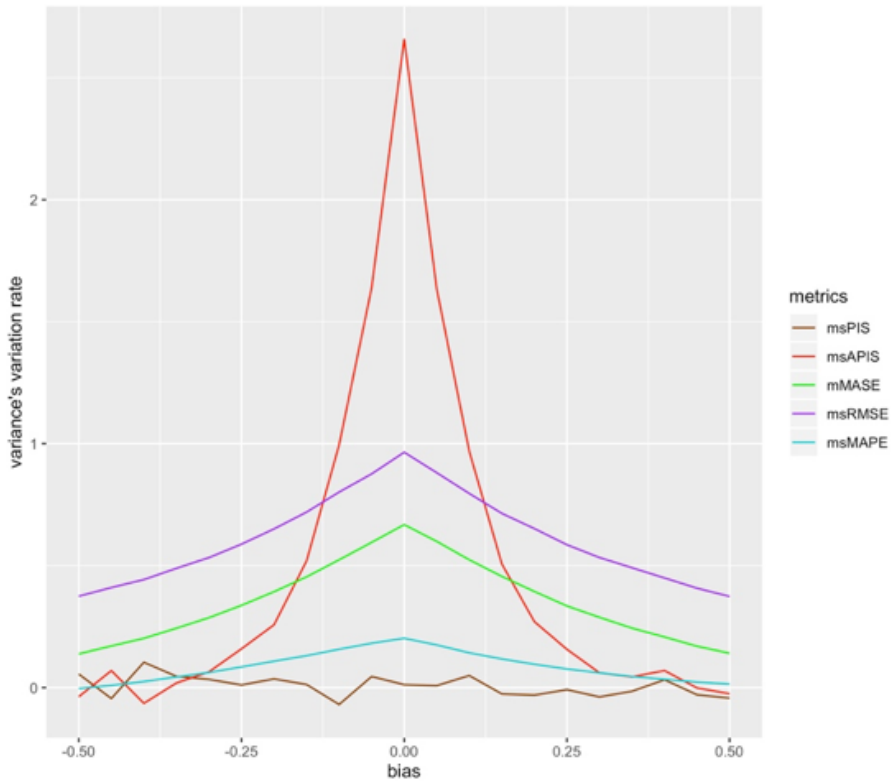

Figure 4.6 Standard deviation variation rate in function of bias

Figure 4.6 presents the variation rate of standard deviation when the bias is changed. It shows how the variation rate in the direction of standard deviation is influenced by a change in bias. $msAPIS$ is strongly influenced by a change in bias and the variation rate in the standard deviation direction goes to values close to zero for high absolute values of bias. These results allow one to conclude that the most sensitive metric to standard deviation would be $msRMSE$ in the presence of high bias. $msAPIS$ is the most sensitive metric to standard deviation in cases in which the absolute bias is less than 5% of the in-sample mean. It is also the least stable metric to variation of standard deviation in the presence of high bias. $msMAPE$ is the least sensitive metric, which has a variation rate close to zero for any bias value. $msPIS$ is not affected by standard deviation since averaging the different series errors cancels the negative and positive errors.

4.4.4.2 Bias in function of standard deviation

The same process applied in section 4.4.4.1 can be applied with the bias in function of standard deviation.

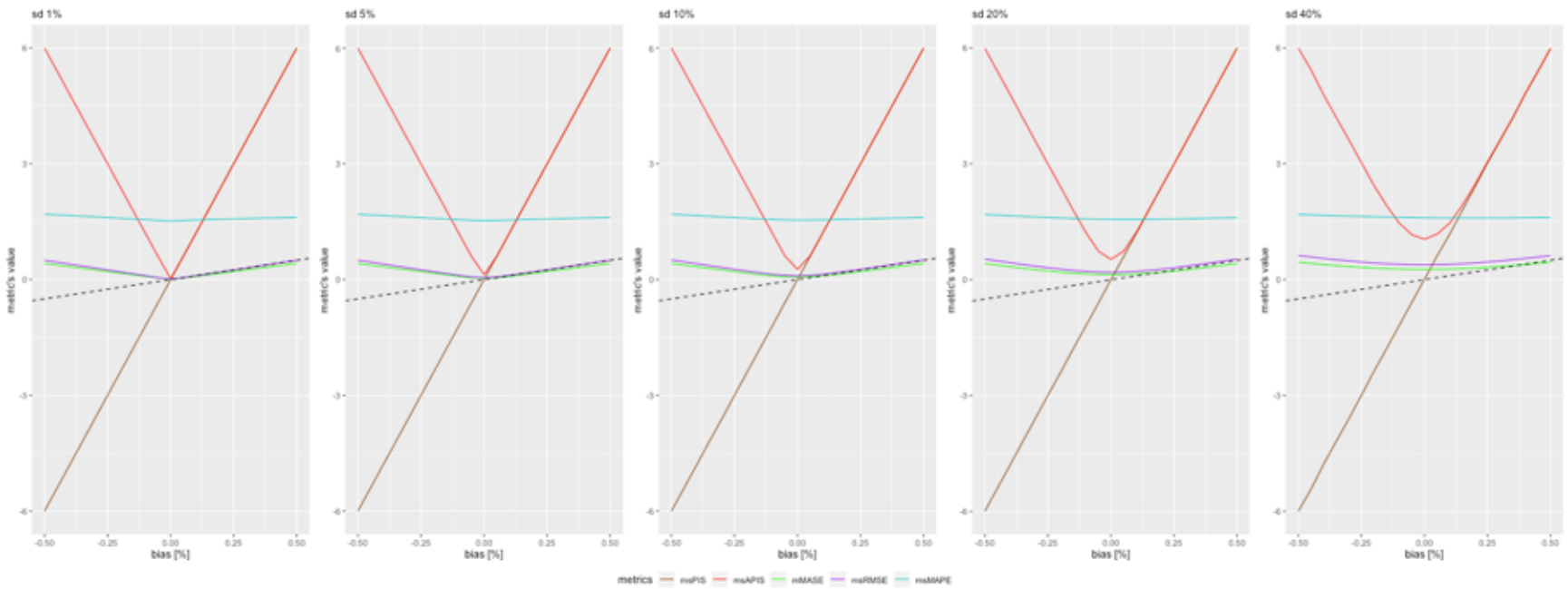

Figure 4.7 impact of change in standard deviation on the different performance metrics

Figure 4.7 shows the impact of a change in standard deviation on the shape of the metric's value in function of bias. Here, the variation rate change around the null bias. This is the case for all series, but the effect is more pronounced for $msAPIS$. Indeed, the variation seems to flatten and the flattening seems to increase with the standard deviation. This could be explained by the fact that a higher standard deviation means a higher metric value. So, the origin in the case of null bias is higher in the presence of high standard deviation, and a small increase in the bias is insignificant compared to the standard deviation.

Let us evaluate the change in average absolute variation rate of the bias in function of the standard deviation.

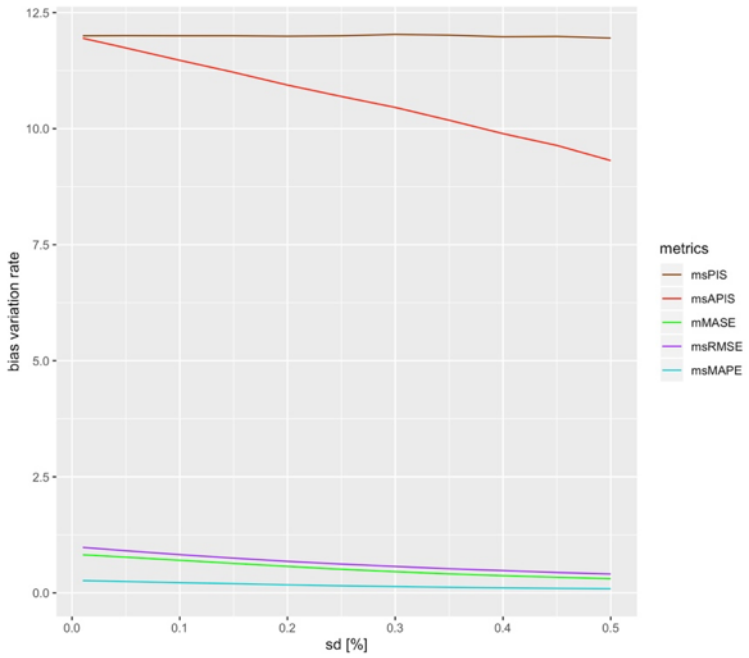

Figure 4.8 Bias variation rate in function of standard deviation

Figure 4.8 presents the variation rate of the bias when the standard deviation is changed. $msAPIS$ is the second most sensitive metric after $msPIS$, which is constant for all standard deviation values. All the other ones are close together, with $msRMSE$ slightly more sensitive than the others and $msMAPE$ is the least sensitive metric after $msPIS$.

4.4.5 Reliability of the metrics

For every configuration, reliability has been measured in function of the number of series averaged performance and the number of forecast horizons. To simplify visualization, only extreme values of forecast horizons were kept: $h = 1$ and $h = 23$. The + and – configurations,

as presented in Table 4.1, were set to be of one order of magnitude different. The different values used are presented in Table 4.4. In each configuration, 25 different models were trained with a difference of $\Delta\sigma$ or $\Delta\mu$ for their error distribution parameter.

4.4.5.1 Reliability to a change in standard deviation

Three different cases can be distinguished when changing the standard deviation of the models. The first case is the one with all the models of the same order of magnitude so σ, μ and $\Delta\sigma$ in – or + configuration. The second is when the bias is of a high magnitude and the third is when the bias has a small order of magnitude. Since $msPIS$ is invariant to standard deviation, it was removed from the following figures.

4.4.5.1.1 Same order of magnitude The figures 4.9 and 4.10 present Spearman's rank correlation and nDCG in function of the number of series performance averaged. The two different types of lines present the number of forecast horizons used. The solid line presents the case in which a single point forecast is evaluated. The dashed line is for cases with 23 forecast horizons.

Figure 4.9 and Figure 4.10 both show the superiority of msRMSE in ranking the models in the correct order. Even though the correlation of mMASE and msRMSE show that their rankings are both going in the same direction as the real ranking, nDCG clearly distinguishes both metrics, with msRMSE converging more quickly to a perfect ranking. sMAPE presents some interesting properties that will be discussed in section 4.5.

4.4.5.1.2 High bias This section presents the results for a configuration with a high (+) bias.

In this case, Figure 4.12 shows that in the presence of high bias compared to standard deviation, no metric can detect changes in standard deviation no matter how many observations are used, except for $msRMSE$. In both configurations, $msRMSE$ is the most reliable metric. Results in Figure 4.11 correspond to what was found in section 4.4.4.1 where most metric sensitivity to standard deviation decreased in the presence of a high bias. The least impacted

Table 4.4 Values of each parameter for all of the different configurations

Cases	μ		σ		$\Delta\sigma$		$\Delta\mu$	
	+	-	+	-	+	-	+	-
Fixed Bias	2%	0.1%	1%	0.01%	0.1%	0.01%	0	
Fixed Standard deviation	1%	0.01%	2%	0.1%	0		0.1%	0.01%

Figure 4.9 Spearman Rank correlation and nDCG for $(+++)$ configuration

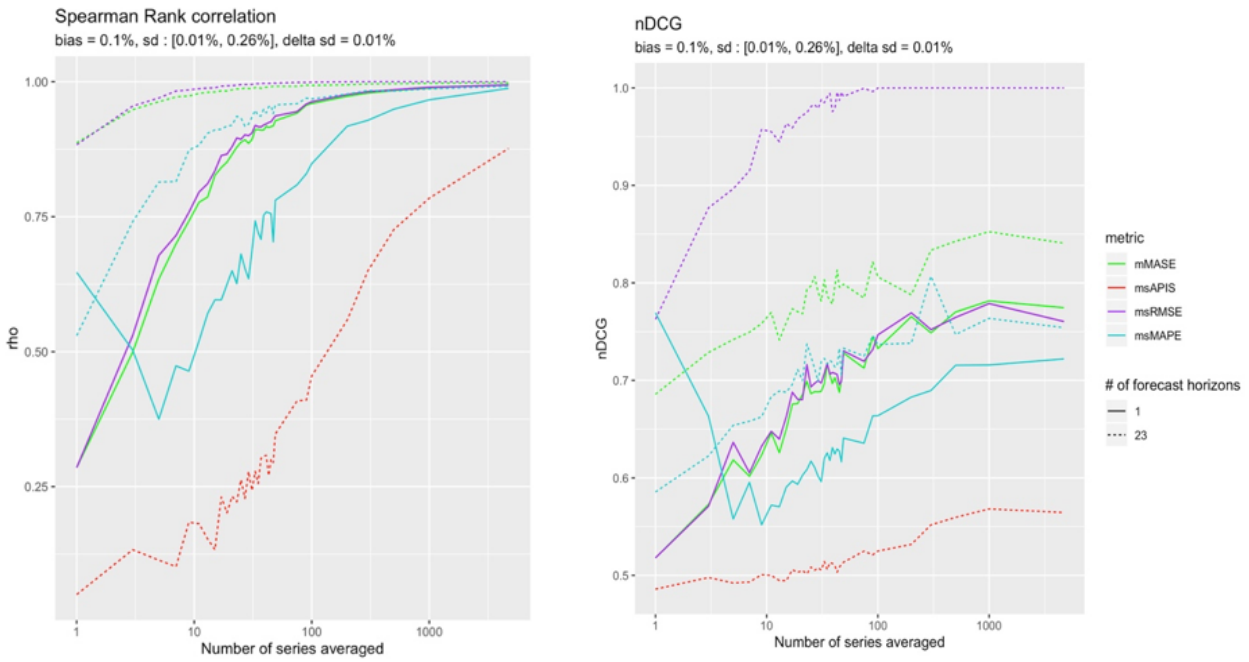

Figure 4.10 Spearman Rank correlation and nDCG for $(---)$ configuration

metrics by bias were $msRMSE$ and $mMASE$. Figure 4.6 also shows that $msRMSE$ was slightly more sensitive to standard deviation in presence of a high bias than $mMASE$. This trend might increase when increasing the difference between the two parameters. This could

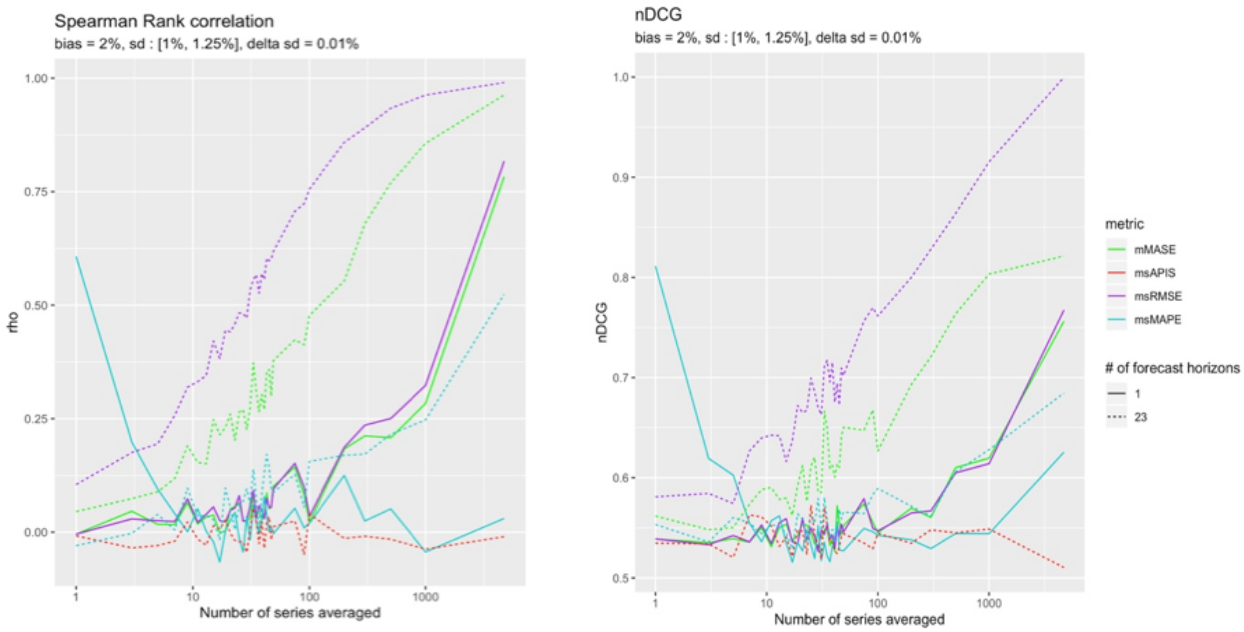

Figure 4.11 Spearman Rank correlation and nDCG for $(++-)$ configuration

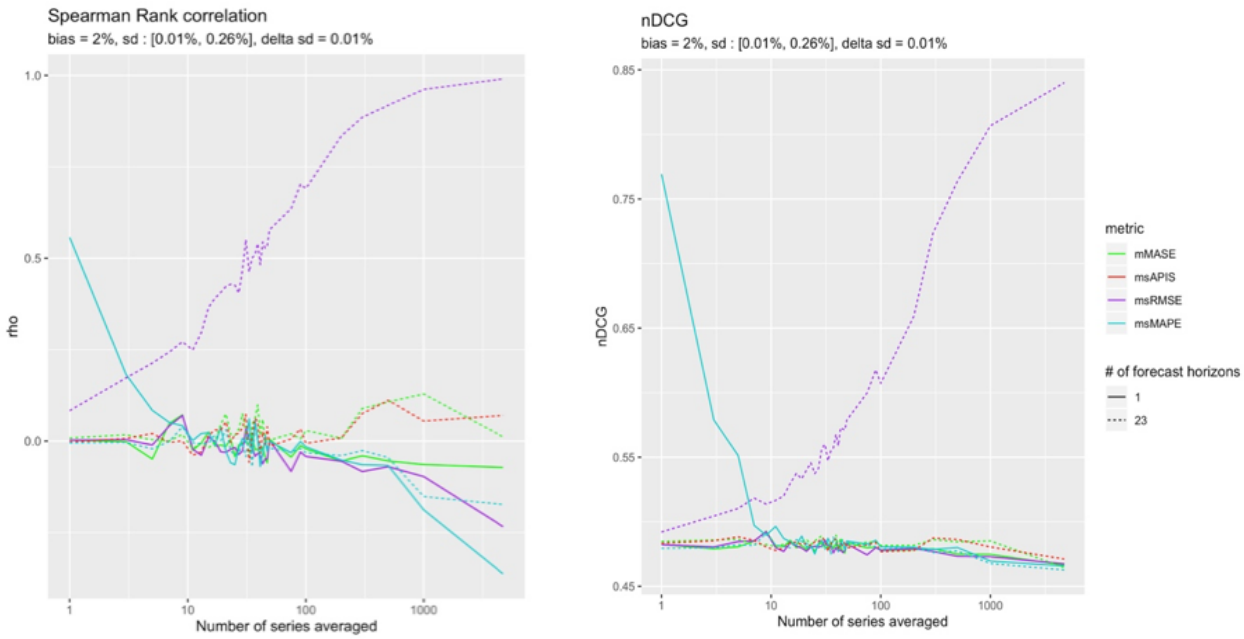

Figure 4.12 Spearman Rank correlation and nDCG for $(-+-)$ configuration

explain the results of Figure 4.12.

4.4.5.1.3 Small bias The remaining cases are those in which a bias is in a (-) configuration. The figures 4.13 and 4.14 show which metric performs better in the quasi-absence of

bias. While the results are tighter, it is still possible to distinguish slightly more reliable results from $mMASE$ in the presence of 23 forecast points. $msRMSE$ is slightly more reliable for a single point forecast.

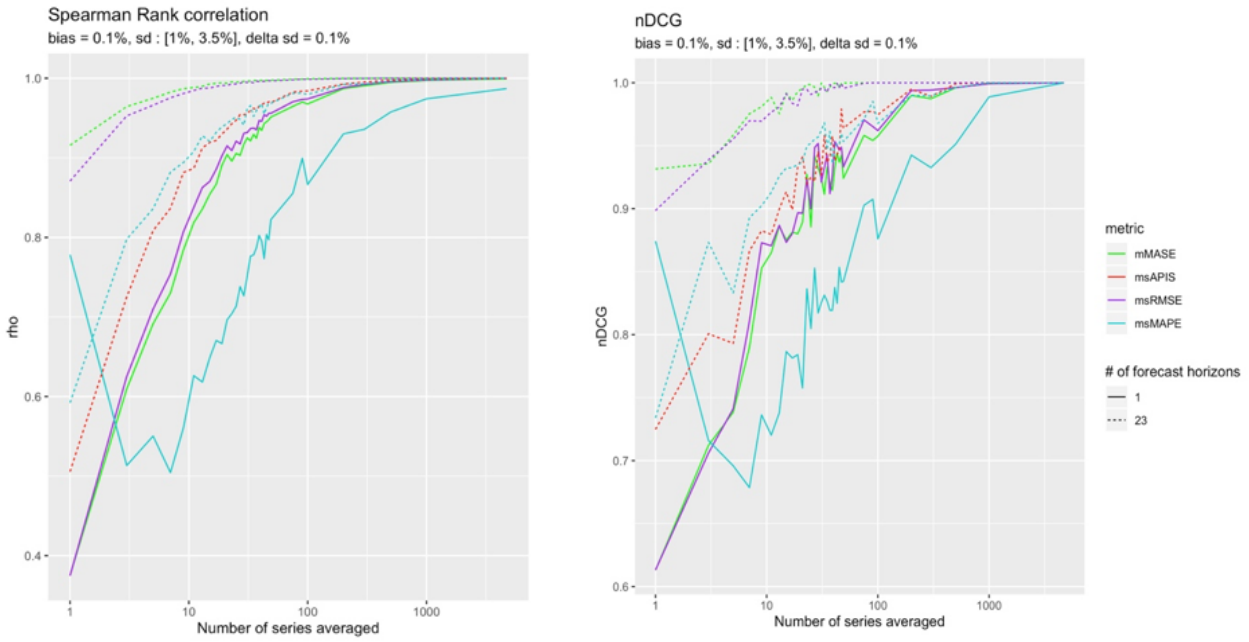

Figure 4.13 Spearman Rank correlation and nDCG for $(+-+)$ configuration

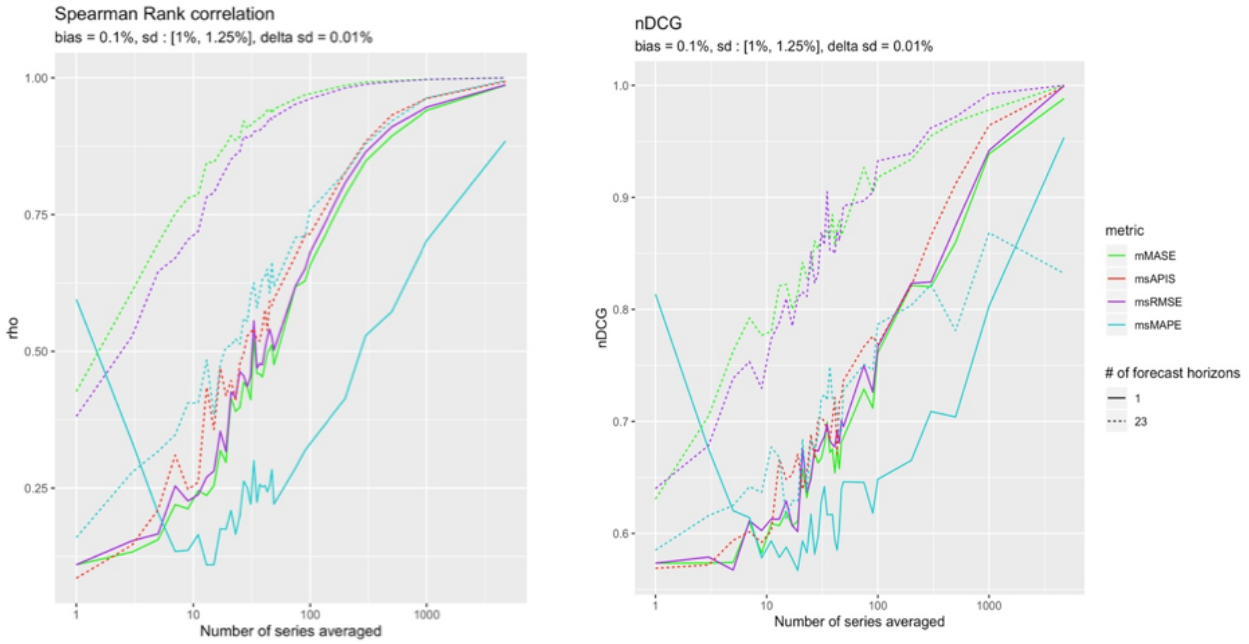

Figure 4.14 Spearman Rank correlation and nDCG for $(+--)$ configuration

4.4.5.2 Reliability to change in bias

When it comes to detecting change in bias, three cases can be distinguished. The first one is configurations with high standard deviation and high bias. The second one is high standard deviation and small bias. The third one is cases with small standard deviation configurations.

4.4.5.2.1 High standard deviation and high bias The two cases with high standard deviation and high bias are presented in the figures 4.15 and 4.16.

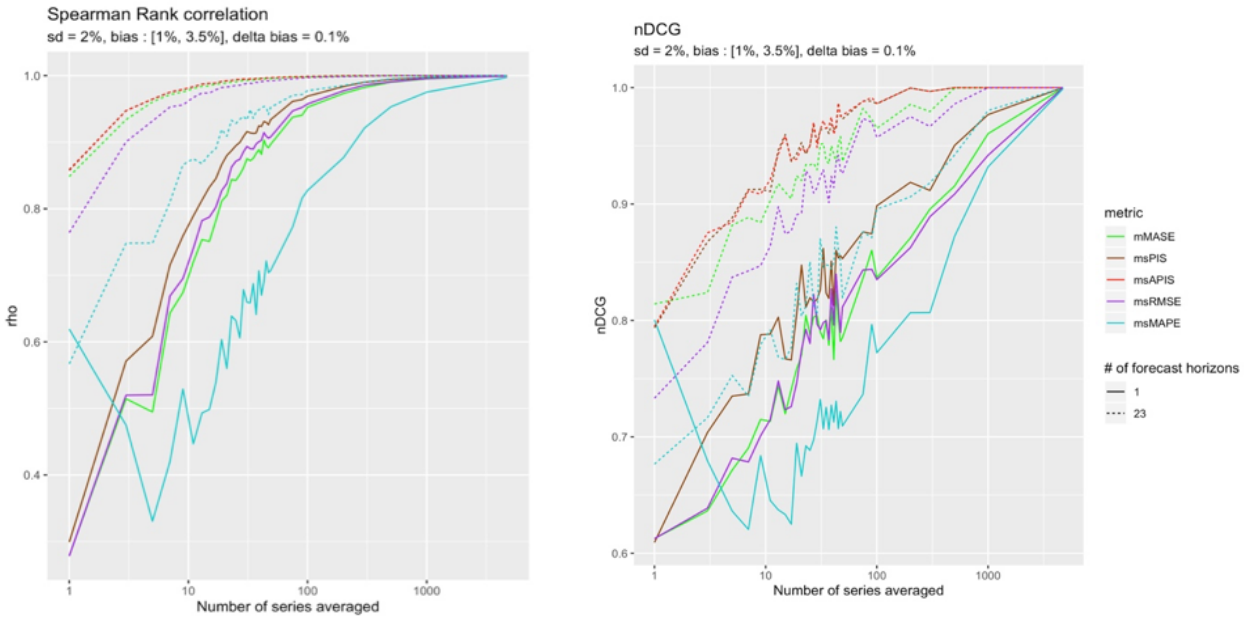

Figure 4.15 Spearman Rank correlation and nDCG for $(+++)$ configuration

As expected, the cumulative metrics $msPIS$ and $msAPIS$ were more reliable than the others in detecting bias. However, $mMASE$ and $msRMSE$ were both able to reach a perfect ranking with all the available observations.

4.4.5.2.2 High standard deviation and low bias This configuration is the only one for which $msPIS$ was the only metric able to converge to a perfect ranking (see Figure 4.17).

This result confirms what was found in section 4.4.4.2, where in the presence of high standard deviation, the sensitivity of most metrics to bias decreases to nearly zero.

4.4.5.2.3 Small standard deviation The final case where standard deviation is in (-) configuration also corroborates with the results in section 4.4.4.2.

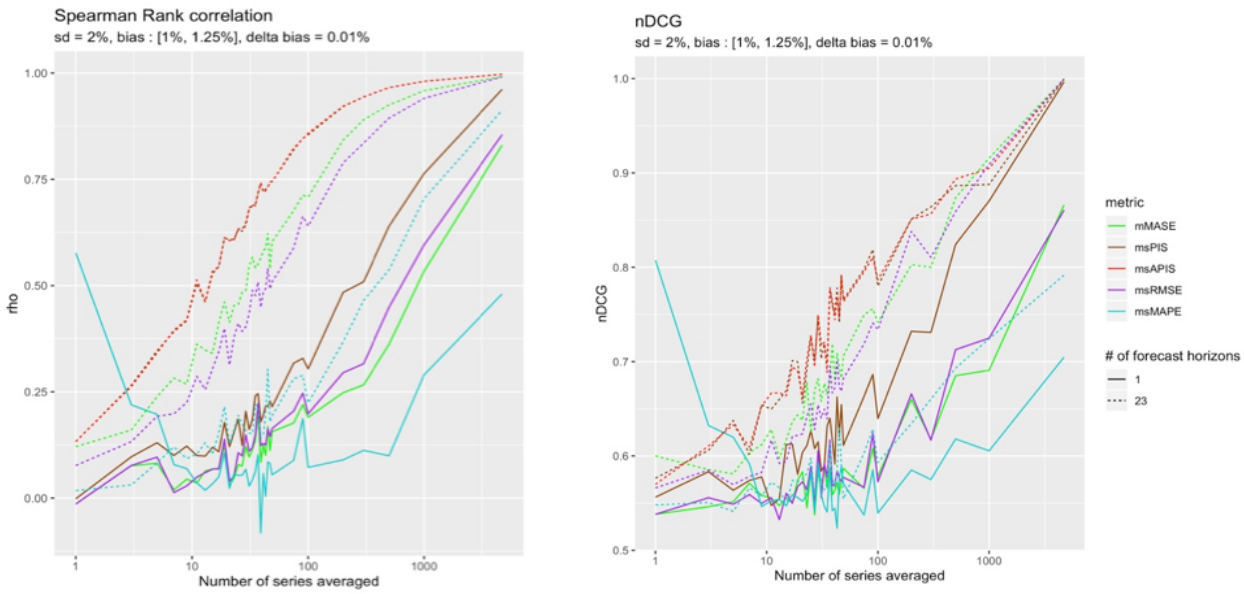

Figure 4.16 Spearman Rank correlation and nDCG for $(+ + -)$ configuration

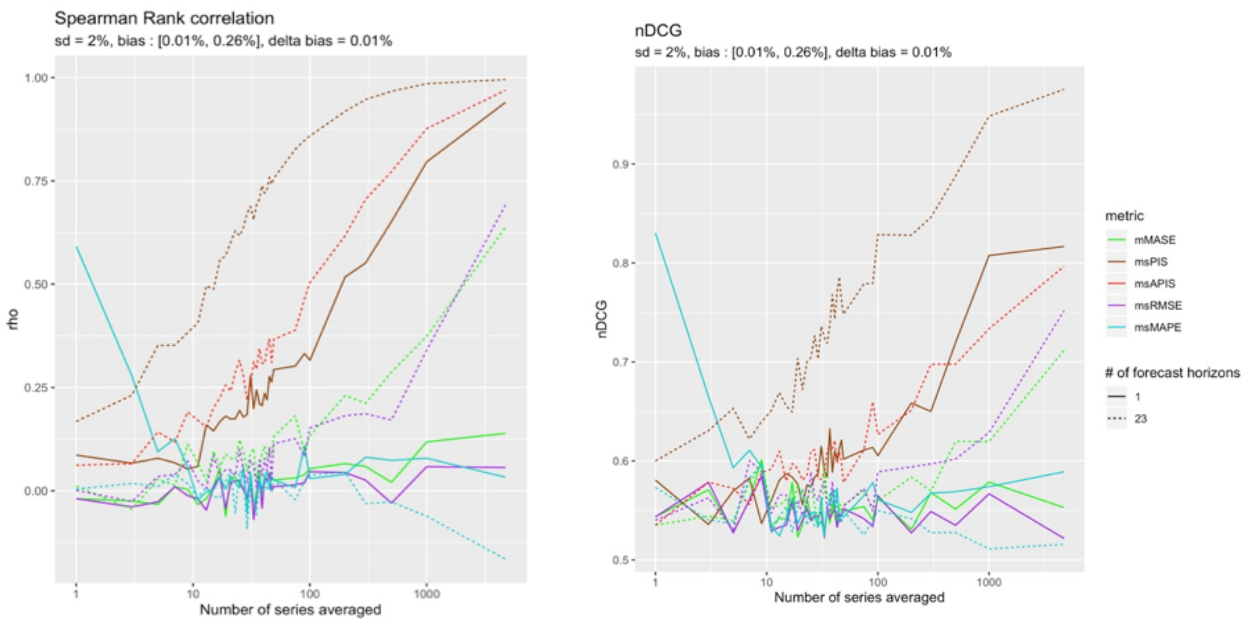

Figure 4.17 Spearman Rank correlation and nDCG for $(+ - -)$ configuration

It has been shown that when standard deviation is small, all metrics have a non-null variation rate for bias. This is what the figures 4.18, 4.19 and 4.20 present.

Indeed, no metric's variation rate for bias is null for small values of standard deviation, but metrics with the most important sensitivity to bias are not superior to the other metrics. This result was found in other cases, such as those in section 4.4.5.1.3, in which $msAPIS$

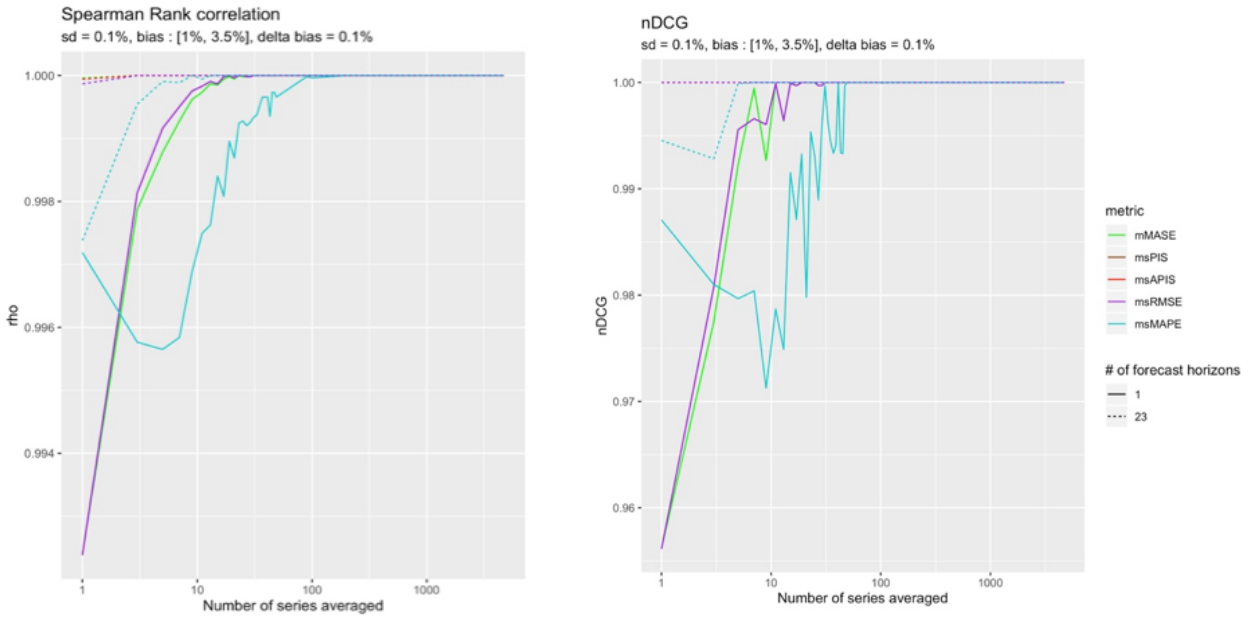

Figure 4.18 Spearman Rank correlation and nDCG for $(- + +)$ configuration

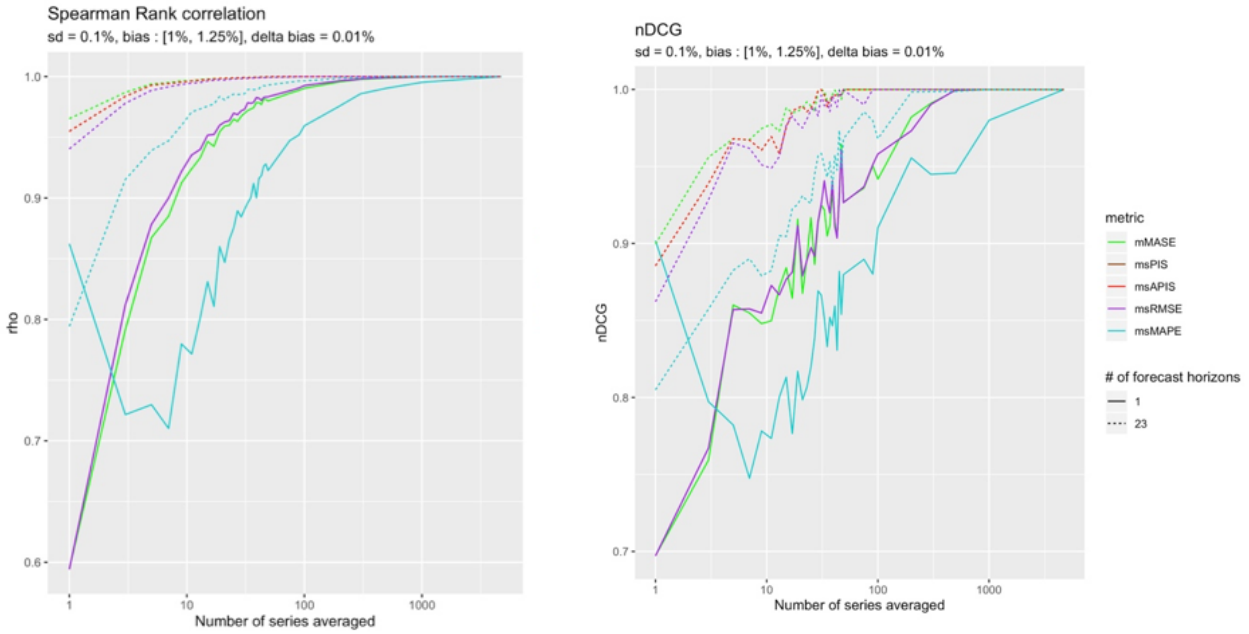

Figure 4.19 Spearman Rank correlation and nDCG for $(- + -)$ configuration

did not perform as well as expected based on the sensitivity results (Figure 4.6). The next section will discuss this and will study further results of a single point forecast for a single series to try to explain the $msMAPE$ results.

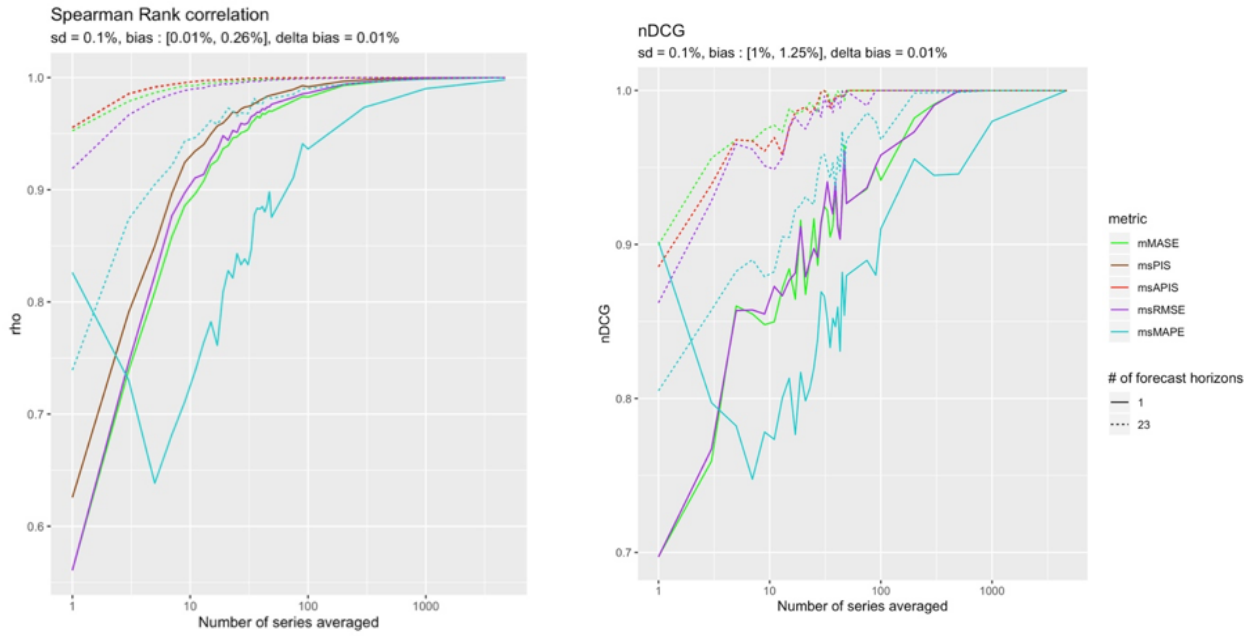

Figure 4.20 Spearman Rank correlation and nDCG for $(- - -)$ configuration

4.5 Result Analysis

Following the reliability results, one would conclude that sensitivity did not have much impact on reliability. However, to measure sensitivity, all 23 forecasts horizons were used with all series. If the sensitivity experiment is rerun with 1, 10 and 100 series instead of thousands, it is possible to see how the amount of series affects sensitivity and reliability (Figure 4.21), which explains reliability results.

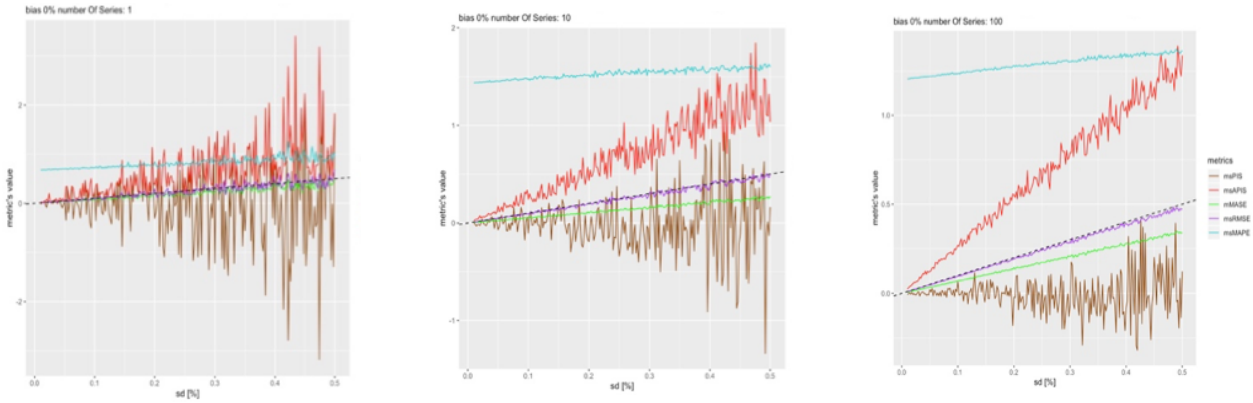

Figure 4.21 Standard deviation sensitivity when averaging results of 1, 10 and 100 series

From the figure 4.21, the observation is that the cumulative metrics are the most affected by the number of series used to average results. That makes sense with previous reliability

results, in which the increase in reliability for $msAPIS$ could only be observed for a high number of averaged series performances. To further study the results of $msMAPE$ for a single series and a point forecast, the level of intermittence allowed in the horizon periods varied to see whether the proportion of zeros in the horizon periods impact the reliability of $msMAPE$. Figure 4.22 shows the reliability of all metrics when the allowable proportion of zeros is less than a threshold. So, when the proportion of 0 allowed in a series is 0, it means that only the series with no zero demand within their horizon periods were kept. On the other hand, if the proportion of 0 demand allowed is 1, it means all series were kept. The average results of all the single series respecting the threshold are represented in the figure 4.22.

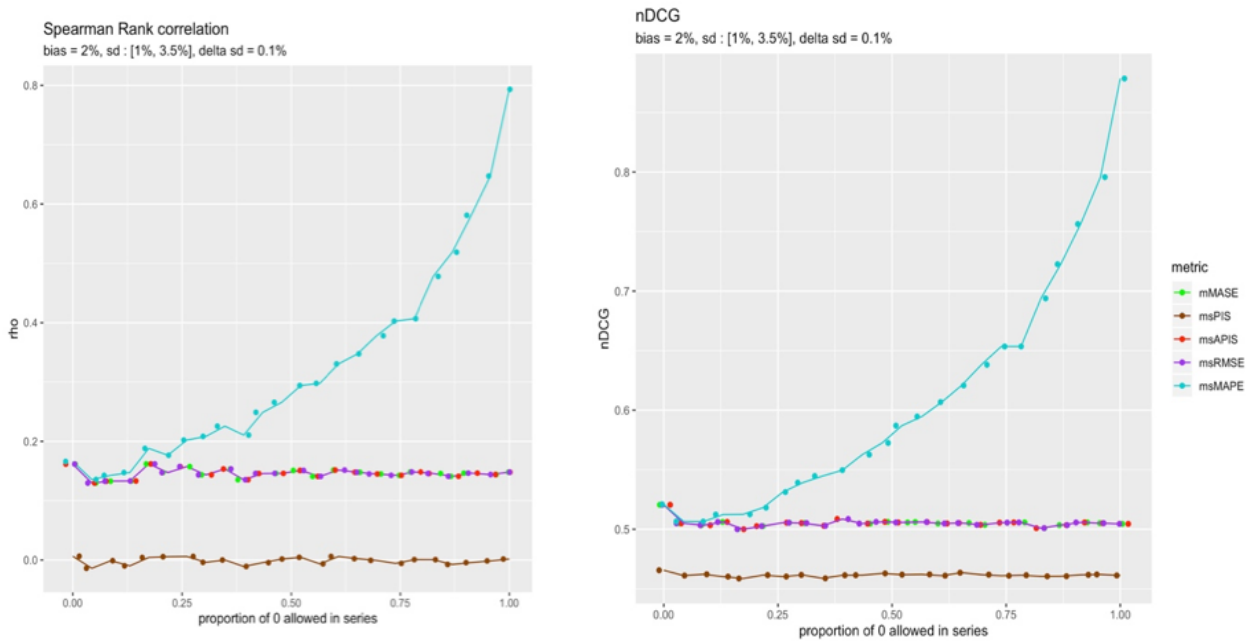

Figure 4.22 Reliability of series in function of the level of intermittence

There seems to be a relationship between the reliability of $msMAPE$ for a single time series point forecast and the level of intermittence in the series. This is probably caused by the fact that near zero demand $msMAPE$ can explode to infinity, making its sensitivity to small errors greater. Finally, Table 4.5 summarizes the results of section 4.4.5.

In the case of a single time series point forecast, the conclusions in Table 4.5 only hold in the presence of a series with a high level of intermittent demand. Other results show that the best performance metrics are $msPIS$ to detect bias and $msRMSE$ to detect standard deviation. So, to detect both the bias and the standard deviation, one must first select models of the same order of magnitude according to the absolute value of $msPIS$. This ensures to keep

Table 4.5 Best choice of metric for different cases based on the reliability results

	Number of series					
Point forecast	**1**		**<1000**		**>1000**	
	sigma	**mu**	**sigma**	**mu**	**sigma**	**mu**
Single	msMAPE	msMAPE	msRMSE	msPIS	mMASE/msRMSE	msPIS
Multiple	msRMSE	msPIS	msRMSE	msPIS	msRMSE	msPIS

the models of minimal bias without consideration to standard deviation. Next, the selected models need to be ranked according to $msRMSE$. This strategy ensures that the models of minimal bias are kept, and then ranked according to their standard deviation.

4.6 Conclusion

The goal of this paper was to present a new methodology to assess the precision and reliability of performance metrics. Fictitious forecasting models were defined as the addition of a noise of a known distribution to the actual values of the series. Given that the error distribution of the models was known, it was possible to estimate the sensitivity of metrics to changes in bias and standard deviation of the fictitious models. It was also possible to rank the models based on their error distribution, allowing the reliability of performance metrics to be studied in different cases. It is to the best of our knowledge a first attempt at quantifying the sensitivity of performance metrics. Sensitivity is highly influenced by the number of points used to average the performance. Results have shown that, with thousands of points for the average, $msAPIS$ was the most sensitive metric in most cases, followed by $msRMSE$ and $mMASE$, while the least sensitive metric was $msMAPE$. This result contrasts with previous beliefs that $MASE$ should be preferred because of its mathematical properties. The reliability results showed that, in most cases, $msRMSE$ was the most reliable metric, followed by $mMASE$. The exception is for cases where the models' differences were due to a change in bias. In those cases, cumulative metrics like $msPIS$ and $msAPIS$ were more reliable. A surprising result was the ability of $msMAPE$ to rank single point forecasting models for a single time series with much more reliability than the other metrics. This result is related to the level of intermittence of the time series. Removing intermittent time series with a high proportion of zeros from the dataset brings the reliability of $msMAPE$ closer to other metrics' reliability.

Thus, the results offer a new perspective on performance metrics, where the proposed methodology has allowed to figure some metrics were more reliable to changes in bias or in variance. Therefore, we propose a strategy to select the best forecasting model by first selecting mod-

els with the same order of magnitude of the absolute value of $msPIS$ and then ranking the selected models based on $msRMSE$.

Future work could study nDCG with relevance in function of both bias and standard deviation to verify how much more reliable selection techniques are when using this last strategy in comparison to mean rank methods.

Acknowledgments

The authors express their gratitude to their industrial partner, Logistik Unicorp for their collaboration in this project and for the supply of data. The authors also thank MITACS (IT 12058) for the financial support.

CHAPITRE 5 ARTICLE 2: MULTIPLE FORECASTING MODELS SELECTION IN THE CONTEXT OF INVENTORY MANAGEMENT

St-Aubin, P., Agard, B.

Abstract *- Forecast accuracy does not necessarily lead to better inventory performance. Therefore, it becomes important to evaluate forecasting models under the conditions they will be used. This paper presents a multi-forecasting models selection methodology in an inventory management context. This is done through a data-driven simulation to evaluate the performances of different forecasting models given a fixed inventory policy. We also propose new multiple forecasting models selection techniques and evaluate the lift in inventory performance induced by the selection of forecasting models based on simulation results versus traditional accuracy metrics. The lift induced by the selection of multiple models compared to the selection of a single model is also evaluated.*

Keywords: Model Selection, Forecasting, Simulation, Inventory management, Inventory control

5.1 Introduction

The driving factor that has led research on demand forecasting so far has been the development of more accurate Forecasting Models (FM). However, this factor does not always seem to improve inventory management performance. Indeed, many research papers have reached different conclusions about the impact of improving forecasting accuracy on inventory performance (Teunter and Duncan, 2009), (Sanders and Graman, 2009). For this reason, before selecting a FM, it is important to test it under the conditions it will be used and to evaluate it compared to the decision variables it impacts.

We propose a simulation method based on historical data to evaluate FM according to their inventory performance. Additionally, this method makes it possible to assess the relevance of using several FM, either by items, by periods, or both.

Until now, work on the selection of models has not explored the possibility of keeping several FM depending on their contextual performance. To this end, we also suggest three new multi-model selection methods that are validated by cross-validation. The impact of simulating performance as well as selecting multiple models is explored through 3 experiments. These experiments make it possible to estimate the lift induced by the simulation of inventory performance, then by the selection of multiple models.

The next sections present, in order, previous and related work 5.2, the simulation and selection process methodology 5.3, the experiments 5.4 and the results 5.5.

5.2 Previous Work

The idea of measuring forecast performances based on decision variables directly started with Gardner (1990). He used trade-off curves to characterize forecasting techniques in terms of delay of service and inventory investment. His results showed forecasts had a great impact on investment and service. Later, (Sanders and Graman, 2009) and (Syntetos et al., 2010a) tried to quantify the relation between accuracy performance and cost. They both found that improving the accuracy reduced the costs. Sanders and Graman (2009) noted in their experiment that error's standard deviation is linearly correlated to the cost but that over a certain threshold bias seemed to have an exponential impact on cost. Syntetos et al. (2010a) used empirical results gathered from a simulation based on historical data to show that a reduction of 1% in MAPE or sMAPE could result in a reduction of 15%-20% of stock and increase cycle service level and fill rate by 1%. However, those results were obtained on items with smooth and continuous demand. According to other findings, those relations between accuracy and inventory performance do not always hold in an intermittent demand context.

Teunter and Duncan (2009) showed in an intermittent demand context that forecasting zeros only can result in a lower RMSE and MAD than other forecasting methods including Croston's method and SBA. To overcome this challenge, they simulated the forecasting method to drive an order-up-to level (T,S) policy. Their results showed the Croston's based method and bootstrap method are superior when measured on inventory performance, i.e., service and stock level indicators. It showed that optimizing accuracy metrics does not necessary lead to lower costs or improved service.

Even though Hyndman and Koehler (2006) have solved the problem of measuring forecast accuracy in intermittent demand context, other authors have used simulation to assess inventory performance of forecasting methods. For example, Syntetos et al. (2010b), Kourentzes (2013), Van Wingerden et al. (2014), and Solis (2015) each assessed the performance of FM by simulating service and stock performance given an Inventory Performance (IP). Following the same idea, Babai et al. (2012) and do Rego and de Mesquita (2015) both studied the impact of temporal aggregation on inventory performance. Similarly and (Barrow and Kourentzes, 2016), studied the impact of forecast combinations on safety stock. Babai et al. (2009), compared the performance of a dynamic versus static reorder point policy. They showed their results were highly sensitive to the forecasts performances. This confirms the need to test FM in comparison to the decision variables it impacts.

More recently, Kourentzes et al. (2019) proposed a simulation optimization approach to optimize forecasting model's parameters directly based on a cost function derived from cycle service level. His results showed that models optimized on MSE were more accurate but a model's bias was minimal when the FM was optimized on the cost function. They also affirm that model selection and combination remains an open question and suggest similar methods could be used to explore possible solutions. This is what is done in this paper. We propose a simulation framework to select multiple models based on different selection methods and different cost functions.

So far the work done on this subject provides evidence that selecting a model for each individual time series is beneficial under smooth demand conditions (Tashman and Kruk, 1996), (Hyndman et al., 2002). However, this conclusion does not seem to hold under intermittent demand conditions. Kourentzes (2014) showed single selection provided slightly more accurate results than multi-selection. Since results according to smooth demand and intermittent demand do not agree, it would be of interest to study the impact of multi-selection in the presence of both intermittent and smooth demand. Moreover, the impact of multi-selection has not been studied according to inventory performance.

Most previous work focuses on either smooth demand or intermittent demand. The dataset in this paper, however, is composed of both intermittent and smooth demand with some series presenting seasonality. It is especially interesting to test for multiple forecasting model selection in this case as some FM are specifically designed for intermittent demand while others perform better in presence of seasonality and smooth demand.

The next section explains the details of the methodology we propose to simulate inventory management and the model selection process.

5.3 Methodology

This section presents the proposed methodology to select a FM in an inventory control context. The idea is to simulate, from historical data, the performance of a set of FM given a selected IP. Then, we use Cross-Validation (CV) to validate the performance of dynamic or hybrid configurations using multiple different FM on different items, different periods, or both.

To apply the methodology, one requires the historical data on each item to forecast and the lead time of each item to reconstruct the events. To initialize the simulation one needs to define parameters to partition the data: the initial date for the simulation (t_0), the actual date which is increased at every time step and is initialized at t_0 (t), the last date of the

simulation (t_v), and the last date of the validation set (T). Figure 5.2 presents a visualization of the parameters. Another parameter to be set is the initial stock for each item. To consider the different scale of demand for each item, the initial stock is considered as a proportion (ρ) of the safety stock (ss).

Figure 5.1 presents a flow chart of the simulation. Its components as well as the parameters of the simulation are described in the following subsections.

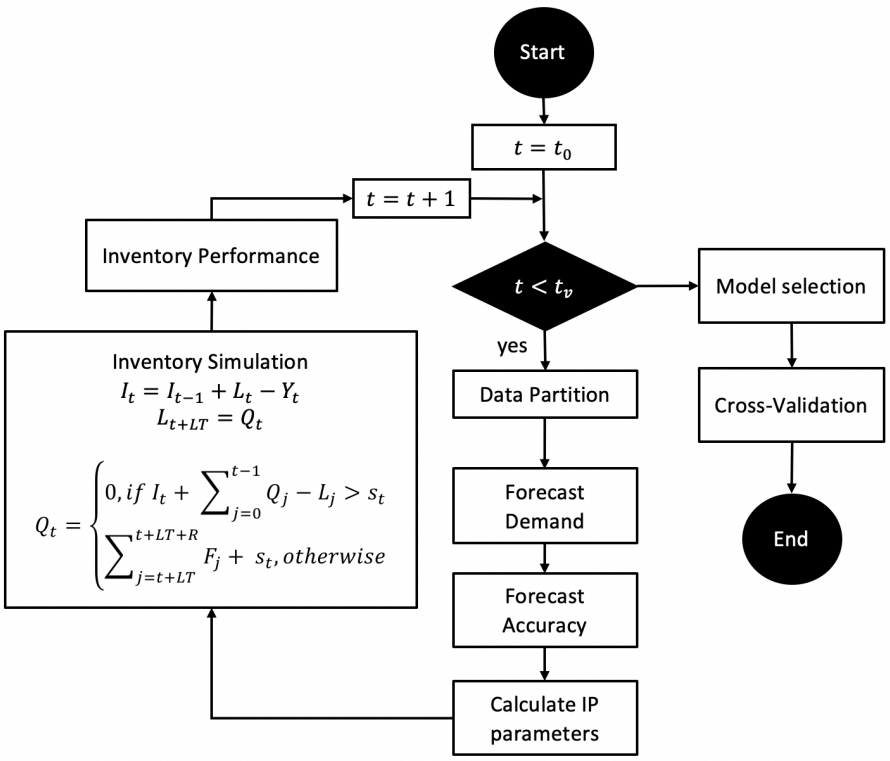

Figure 5.1 Simulation flow chart

5.3.1 Data Partition

The data partition is done dynamically following the parameter t. This parameter represents the actual period ending the training set. The partitioning is done in three sets as presented in figure 5.2.

The training set is used to calibrate the FM. It is changed every iteration with the increment of t. However, note that the minimum value for t is t_0 which sets the minimum number of periods to be used in the training set.

The test set begins at period t and ends at period t_v. It is used to evaluate the performances of the FM in an inventory management context. The validation set starts at period $t = t_v$

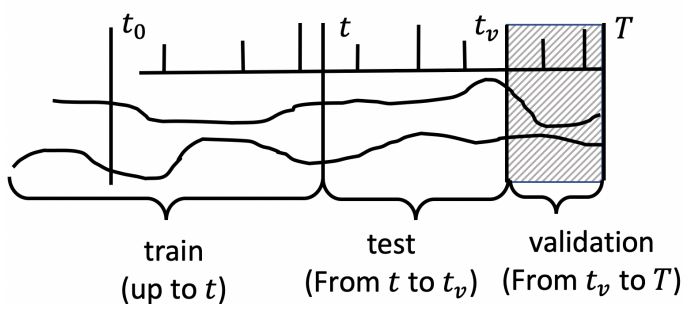

Figure 5.2 Data Partition

to the end of the last period T used for the simulation. It is used to confirm the generality of the performance of the selected FM. For example, a combination of different FM could yield better inventory performance than the selection of a single one. Therefore, the selection of a combination or a sequence of FM acts as the selection of a parameter in a model. To confirm that the improved results obtained with a given model configuration are not due to overfitting the test set, the performance of the configuration should also yield better results on the validation set.

5.3.2 Forecast Demand

Once the data partition is done, the parameters of the chosen FM for the simulation are selected to minimize either the Akaike Information Criterion (AIC), if available, or the Mean Squared Error (MSE) in-sample (on the training set). For every new period (t), the parameters of the models are reevaluated as Syntetos et al. (2010b) and Babai et al. (2009) showed it improved results compared to static parameters.

5.3.3 Forecast Accuracy

The accuracy of the forecasts is measured using the MSE. They are later used to calculate the IP parameters. The MSE is estimated in-sample (on the training set).

5.3.4 Calculate Inventory Policy Parameters

Many different methods can be used to calculate IP parameters. In this paper, the IP is fixed as a (s, Q) policy and the reorder point is evaluated by:

$$s = \sum_{i=t}^{t+LT+R} F_i + \Phi^{-1}(TSL)\sqrt{MSE_t \cdot (LT + R)} \tag{5.1}$$

The first part of the equation (F_i) is the forecasted demand for period i. The second part of the equation represents the safety stock (ss) where Φ is the cumulative distribution of a normal distribution in-sample at period t_0. TSL is the Target Service Level, MSE_t is the in-sample Mean Square Error evaluated with data from t_0 to t, LT is the Lead-Time, and R is the number of periods covered by an order. This safety stock calculation method follows the normal approximation as presented in (Axsäter, 2006). Note that given the actual simulation framework, different calculation methods could be used to estimate the best set of parameters. The order quantity (Q) depends on the stock level and the forecasted values. Its formulation is explained in the following section.

5.3.5 Inventory simulation

The inventory simulation uses the real demand data from period t_0 to T with the corresponding forecasted demand values of the different FM to rebuild, according to the IP, the stock levels and the sequence of orders and deliveries.

To begin, the stock levels at t_0 (I_{t_0}) are initialized as a proportion of the safety stock (ss): $I_{t_0} = \rho \cdot ss$. Where ρ is the proportion parameter of the safety stock ss.

Then, the inventory level (I_t) at the end of the period is obtained by:

$$I_t = I_{t-1} + L_t - Y_t \tag{5.2}$$

Where L_t is the quantity to receive on period t, and Y_t the demand during the period.

Let us define the inventory level with the opened orders (orders passed but not yet delivered) as $IO_t = I_t + \sum_{j=1}^{t-1}(Q_j - L_j)$ The order quantity Q is given by :

$$Q_t = \begin{cases} 0, \text{ if } IO_t > s \\ \sum_{j=t+LT}^{t+LT+R} F_j + s - IO_t, \text{ otherwise.} \end{cases} \tag{5.3}$$

Where F_j is the demand forecast for period j.

Since the opened orders are taken into account for the reorder decision, several Purchase Orders with different due dates can be opened in simultaneously, therefore allowing R to be smaller than LT. Negative stock can also occur. That would translate as back orders.

Finally, the deliveries are set as a time lag of Q_t by LT periods.

$$L_{t+LT} = Q_t \tag{5.4}$$

5.3.6 Inventory performance

The inventory performance of the FM at each period is based on the cost function Z. It evaluates the cost of the Realized Service Level (RSL) for the item i at period t. The cost function is defined as follows:

$$Z_{i,t} = Q_{i,t} \cdot c^{I}_{i,t} + |I_{i,t} - TSL \cdot Y_{i,t}| \cdot \left[\frac{c^{BO}_{i,t}}{2} \Big(1 - sgn(I_{i,t} - TSL \cdot Y_{i,t}) \Big) + \frac{c^{OS}_{i,t}}{2} \Big(1 + sgn(I_{i,t} - TSL \cdot Y_{i,t}) \Big) \right] \tag{5.5}$$

Where $Q_{i,t}$, $I_{i,t}$, $Y_{i,t}$, $c^{I}_{i,t}$, $c^{BO}_{i,t}$ and $c^{OS}_{i,t}$ are the ordered quantity, inventory level, the demand, the item's cost, the back order, and over stock costs for item i at period t respectively. TSL represents the Target Service Level. In this case, the inventory level can be negative as back orders are allowed.

Note that the cost function can still be used in absence of information about cost by attributing a cost of 0 or 1 to the items and of 1 to both back order and over stock costs. In that case, the cost function would simplify to the first term of Z. Different penalties to positive or negative errors can also be applied by increasing one of the costs compared to the other.

Once the performance for the period is evaluated, the period t is increased and the algorithm checks if t has reached t_v which would stop the simulation and begin the model selection and cross-validation of the selection.

5.3.7 Model selection

Different selection methods can be applied and compared on the validation set. With our methodology, the selection methods are based on the inventory performance cost function defined above. Observing it with different levels of aggregation could lead to better global inventory performance by combining multiple FM either at different periods or for different items. We use the term model configuration to designate such a combination or sequence of models.

Single Model Sum Selection (SMSS) Selects a single FM based on the total sum of costs over all items from $i = 1$ to N and all periods from t_v to T : $min \sum_{i=1}^{N} \sum_{j=t_v}^{T} Z_{i,j}$. This way, the model that yields the lowest cost over all items and all periods gets selected. Note that the aggregation operation is changed for the average when used with classical accuracy metrics.

Single Model Ranked Selection (SMRS) Selects a single FM. It takes the sum of the cost over all periods. For all items, the models are ranked according to their cost. The FM with the minimum average rank over all items and periods is chosen. This method accords the same weight to all items and therefore should yield better RSL on the global evaluation with physical inventory.

Multi Model Item Selection (MMIS) Selects different FM for different items or groups of items by taking the sum of costs over all periods. For each item or group of items, the FM that yields minimum costs over all periods is selected.

Multi Model Period Selection (MMPS) Selects different FM at different periods. The idea is similar to the previous one, where the sum of costs over all items for all periods or aggregation of periods is taken. The minimum cost models for each period or aggregation of periods are selected.

Multi Model Item Period Selection (MMIPS) This method combines the last two methods. It selects models for different items or groups of items and different periods or aggregation of periods. To do so, the minimum cost models over the periods or aggregation of periods for each item or group of items are selected. This method can result in the selection of different sequences of models for each item.

Multi-model selection can be useful if different items to forecast present different characteristics such as intermittent demand or seasonality. However, such a detailed selection of models could lead to overfitting the test set used for the simulation. This is why cross-validation is required to confirm the generality of the selection on an independent dataset.

5.3.8 Cross-validation

After applying several selection methods, a definitive model configuration is chosen according to the global performances on the validation set. This is done to avoid the selection of models configuration that overfits the test set. The configuration with the best results on the validation set should yield the best results on future observations as the application of train test cross-validation (Hyndman, 2014) proves this configuration performs well out-of-sample.

The final global evaluation to compare the different FM configurations selected is done according to three metrics:

The Total Cost : $TC = \sum_{i=1}^{N} \sum_{j=t_v}^{T} Z_{i,j}$. It sums the costs over all items and periods. One drawback of this metric is that it weighs high demand and expensive items more than low demand item. Thus, it could mask a poor service level performance for certain groups of products with low demand.

The Stock Ordered : $SO = \sum_{i=1}^{N} \sum_{j=t_v}^{T} Q_{i,j}$. It represents the total quantity of stock ordered.

The average Client Service Level : $\overline{CSL} = \frac{1}{N \cdot T} \sum_{i=1}^{N} \sum_{j=t_v}^{T} min(1, PI_{i,t}/Y_{i,t})$. It represents the realized service level from the point of view of a client. This implies the CSL is between zero and one.

The next section presents an experiment following this methodology to confirm its validity and to compare its results to classical FM selection based on accuracy metrics.

5.4 Experiment

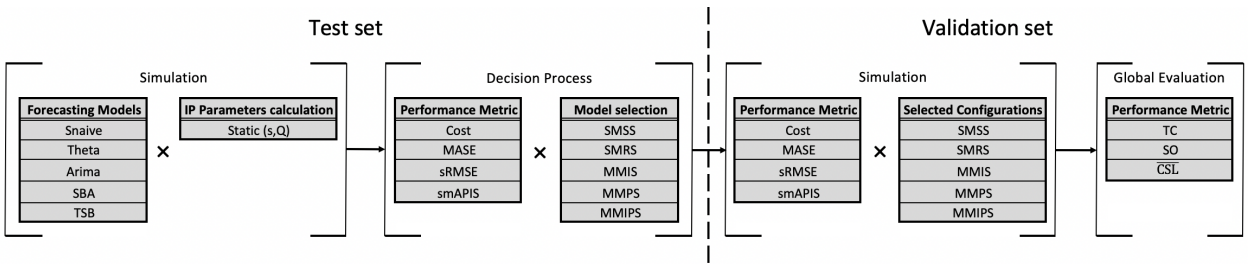

Figure 5.3 Experiment 1: Multi-Model Selection

To test this methodology, we built an experiment using snaive, arima with seasonal component if necessary (Hyndman and Athanasopoulos, 2018), theta (Assimakopoulos and Nikolopoulos, 2000), SBA (Syntetos and Boylan, 2005) and TSB (Teunter et al., 2011) as FM.

Figure 5.3 presents the Experiment 1 setup where first, all FM are simulated with the IP and the calculation method of its parameters on the test set. The range of the order R is set equal to the lead time plus one period ($LT + 1$) as it is common in literature. With the results of each iteration of the simulation, different performance metrics are calculated. Then, a model selection method is applied according to each metric. This results in the selection of 20 models configurations. Those configurations are then used to simulate the validation set. The global performances of all configurations are then compared to make a conclusion on the efficiency of the methodology.

The experiment uses data from a company that supplies uniforms for several different organizations. It contains over 10 000 items to supply. The data contains the demand for each item from 2012 to 2019. The first two years of data are kept for the training so t_0 is set to the first date of 2014. The last complete year of data is kept for the validation set, so t_v is set to the first date of 2018.

The initial stock is set equal to the reorder point, meaning $\rho = 1$ for each item. In the absence of information about the costs, c^I is set to 1, c^{OS} is set to 0.05, and c^{BO} to 0.5 for all items. For the multi-model selection methods, individual items are the aggregation considered for the items multiple selection and a 3 months period aggregation is chosen for the period aggregation.

The accuracy metrics selected are presented in detail in (Hyndman and Koehler, 2006), (Petropoulos and Kourentzes, 2015) and (Kourentzes, 2014) for MASE, sRMSE and sAPIS respectively.

The next section describes the benchmarks and the contribution of the experiments.

5.4.1 Benchmarks and contribution

This experiment serves two purposes:

1. To confirm the validity of the methodology
2. To estimate the lift induced by:
   - 2.1 The selection of FM based on inventory performance
   - 2.2 The selection of multiple FM through the application of new proposed methodology on the selection process

To evaluate these points, two other experiments are carried out. Their details are explained in the following subsections.

5.4.1.1 Confirm the validity of the methodology

To confirm the validity of the methodology, a second experiment is run: Experiment 2. It follows a classical accuracy metric selection method in which a single model is selected according to the SMSS and SMRS, which are a common selection processes in literature (Makridakis and Hibon, 2000).

Figure 5.4 presents the characteristics of Experiment 2. It presents a classical train-test evaluation, where the whole train and test sets are used to train the FM, and the validation set is used to select the most accurate model. The global performance of the most accurate model is then calculated. This last result is used as a benchmark to confirm the validity of the methodology and to evaluate the lifts in 2.1 and 2.2.

If the global performance obtained in Experiment 1 is similar or better than the global performance in Experiment 2, then the methodology is considered valid to select FM.

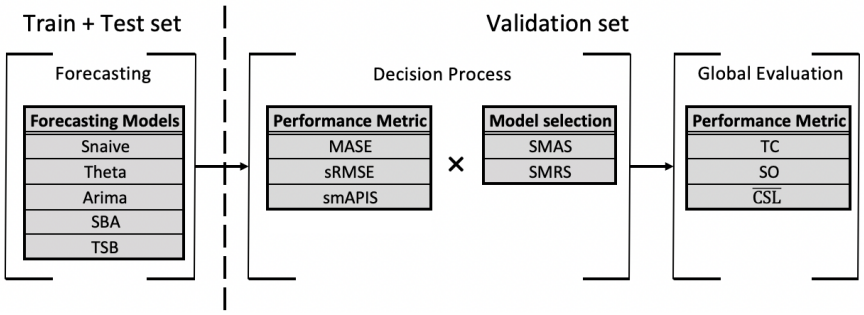

Figure 5.4 Experiment 2: Classical Accuracy Selection

5.4.1.2 Estimate Performance Lift

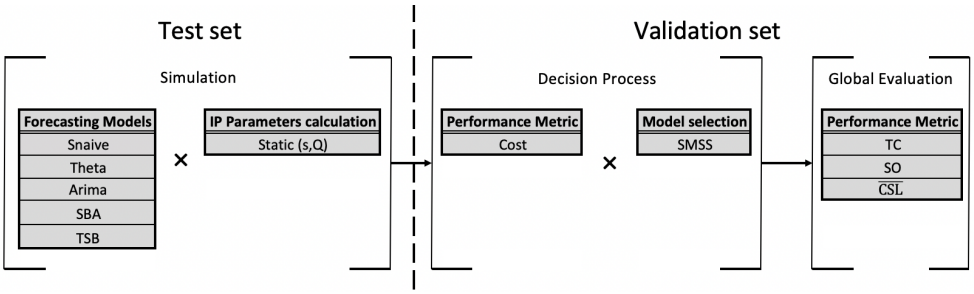

Figure 5.5 Experiment 3: Classical Simulation Selection

The lift induced by the simulation of the inventory performance is evaluated based on the results of the Experiment 3 presented in figure 5.5. In this experiment, again, the selection is made directly on the validation set. The lift is evaluated in two ways. First by comparing the results from the selection based on the cost function in Experiment 3 to those from accuracy metrics in Experiment 2. Second by comparing the global performance of the selection in Experiment 2 to the model with the best global performance.

Finally, point 2.2 is estimated by comparing the global results from Experiment 1 obtained with accuracy metrics to those of Experiment 2. The results from Experiment 1 obtained from the inventory based cost function are compared to those of Experiment 3. Thus, the lift caused by the multi-model selection methodology can be evaluated independently from the performance metric used.

The results of the three experiments are presented, compared and discussed in the following section.

Table 5.1 Experiment 2: SMSS and SMRS results

	SMSS			SMRS		
model	**MASE**	**sRMSE**	**sAPIS**	**MASE**	**sRMSE**	**sAPIS**
SBA	**1.044**	**11.546**	73.400	**2.330**	2.224	**1.012**
TSB	1.071	11.686	47.517	2.916	**2.157**	2.048
arima	1.348	36.554	**46.091**	3.378	3.363	3.100
snaive	1.096	12.530	49.357	3.521	4.519	3.961
theta	1.074	13.179	48.908	2.855	2.737	4.879

5.5 Results

The results of the classical accuracy metrics for single selection methods are presented in table 5.1. The results indicate that SBA seems to be the most accurate model as it has the best performance for most metrics in both SMSS and SMRS.

Table 5.2 Experiment 3: SMSS results

model	**TC** $\cdot 10^6$
SBA	**5.415**
TSB	5.434
arima	5.847
snaive	6.028
theta	6.089

Table 5.2 presents the total cost of Experiment 3. Again, with the cost function, SBA is the model that minimizes the overall costs. To summarize, all of the single selection methods have selected SBA except in the case of SMRS selection with the sRMSE metric, where TSB is selected. Thus, SBA is considered the model selected in both experiments 2 and 3.

Table 5.3 compares the global performances of all configurations. In this case, the results are a little more spread across the global metrics space.

To better represent the results, we plot the results according to the stock ordered and the average CSL on figure 5.6. We decided to plot according to the level of Stock Ordered and average Client Service Level since in this case, the cost information was not obtained from real data but instead set to minimize its impact. The penalties were set to respect the 10% ratio of overstock cost over backorder cost that Babai et al. (2009) and Syntetos et al. (2010b) seem to find realistic.

Table 5.3 Configurations global performance

Selection	model	CSL	SO $\cdot 10^6$	TC $\cdot 10^6$	mRank
Single	SBA	94.7	2.555	5.415	12.833
	TSB	94.9	2.562	5.434	13.000
	arima	**98.0**	2.545	5.847	9.667
	snaive	97.9	2.495	6.028	8.667
	theta	96.4	2.650	6.089	12.667
MMPS	cost	97.5	**2.445**	5.730	6.000
	MASE	94.6	2.508	5.337	9.667
	sRMSE	94.8	2.510	5.348	9.667
	sAPIS	94.6	2.508	5.337	9.667
MMIS	cost	95.5	2.469	**5.065**	**4.000**
	MASE	95.3	2.447	5.341	5.667
	sRMSE	95.4	2.493	5.359	7.667
	sAPIS	94.6	2.548	5.402	12.667
MMIPS	cost	96.1	2.478	5.192	4.333
	MASE	96.3	2.473	5.438	7.333
	sRMSE	95.9	2.489	5.425	8.000
	sAPIS	94.7	2.535	5.407	11.500

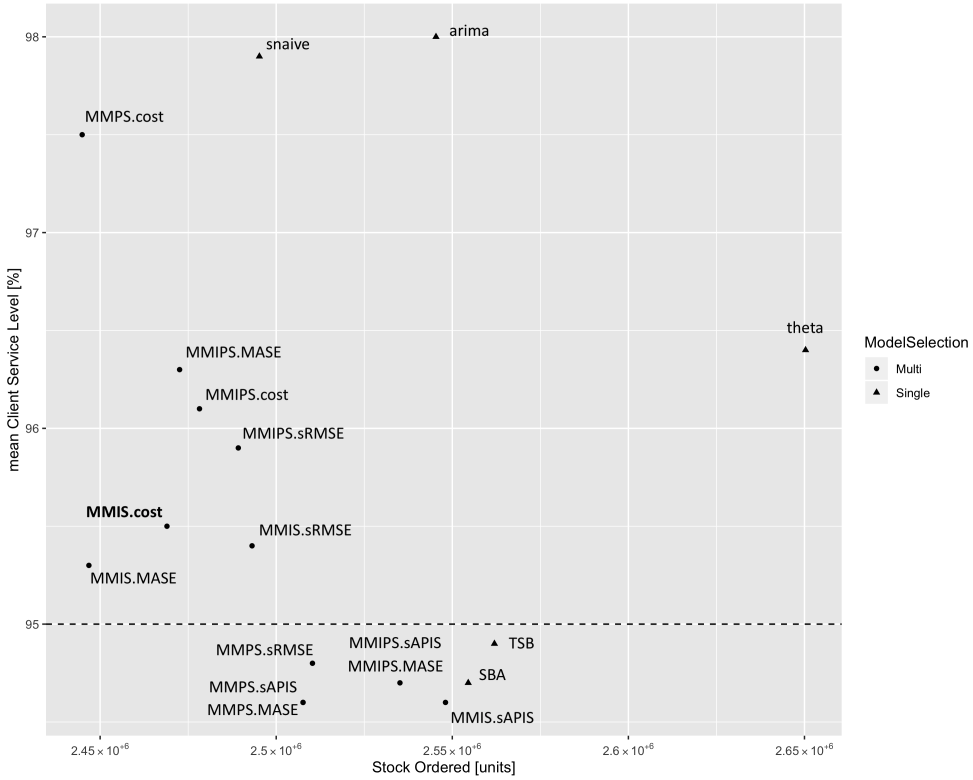

Figure 5.6 All Configurations level of SO compared to the average CSL

One can see on figure 5.6 a clear increase in performance when selecting multiple models over a single one. However, multiple model selections based on sAPIS accuracy metric seem to perform less well compared to the selections using other metrics. This is the same for the MMPS selection where the selection based on accuracy yielded CSL slightly under the target of 95%. However, it could be explained by a poor choice of period aggregation.

5.5.1 Simulation versus Accuracy selection

As presented in table 5.1, single model selection based on accuracy selection would have selected the SBA forecasting method. Now, if we measure the lift of each configuration according to the performance of SBA, we obtain the results presented in table 5.4. The last column of table 5.4 represents the mean lift of all the other columns. Since experiments 2 and 3 selected SBA, it appears the simulation could not improve the selection.

Table 5.4 Mean lift according to selection method

Selection	model	CSL (lift)	SO (lift)	TC (lift)	mean (lift)
Single	SBA	1	1	1	1
	TSB	0.998	1.003	1.004	1.002
	arima	**0.966**	0.996	1.080	1.014
	snaive	0.967	0.977	1.113	1.019
	theta	0.982	1.037	1.124	1.048
MMPS	cost	0.971	**0.957**	1.058	0.995
	MASE	1.001	0.982	0.986	0.990
	sRMSE	0.999	0.983	0.988	0.990
	sAPIS	1.001	0.982	0.986	0.990
MMIS	cost	0.992	0.967	**0.935**	**0.965**
	MASE	0.994	0.958	0.986	0.979
	sRMSE	0.993	0.976	0.990	0.986
	sAPIS	1.001	0.997	0.998	0.999
MMIPS	cost	0.985	0.970	0.959	0.971
	MASE	0.983	0.968	1.004	0.985
	sRMSE	0.987	0.974	1.002	0.988
	sAPIS	1.000	0.992	0.999	0.997

Despite the strong case for this selection, the Total Cost (TC) does not represent the real costs encountered by the company and SBA was not exactly on the target service level. Therefore, one can consider the simulation has allowed the detection of an increase of a little more than 3% on the CSL and a reduction in SO of 2.3% if the snaive model is selected instead of SBA. Additionally, considering the mean rank (mRank) on global performance, snaive and arima both outperform SBA. Based on those factors, one could argue the induced lift by

the simulation is between 0% to 2.8% if taking the average lift over the global performance metrics.

5.5.2 Multi versus single selection

The lift induced by multiple model selection is on average if we take the mean lift on all metrics for all multi-models of 1.4%, with an average increase in CSL of 0.8%, an average reduction in SO of 2.5% and an average reduction in cost of 0.9%. Considering the multi-model selected is the one with the minimum average rank (mRank) on the global performance metric (see table 5.3), the MMIS.cost configuration would be selected. So for the best-case scenario, the lift induced by multiple selection is of 3.5%. With an increase in CSL of 0.8%, a reduction in SO of 3.3% and a reduction in cost of 6.5%.

5.5.3 Impact of cost on lift

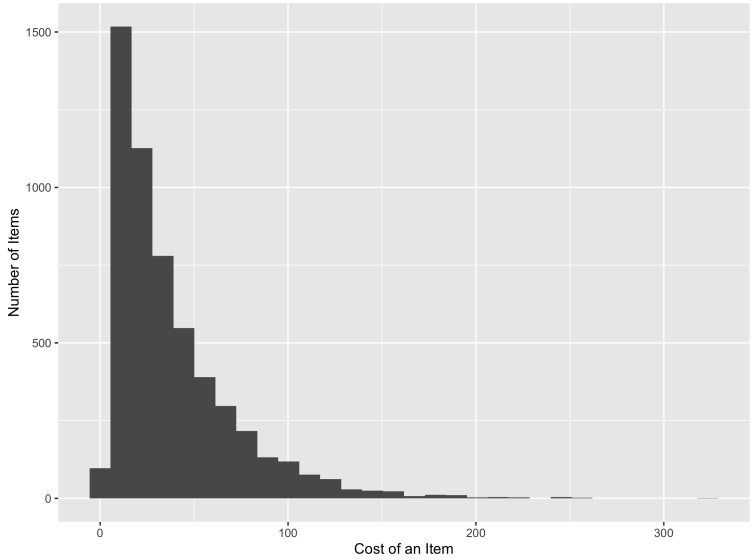

Figure 5.7 Distribution of item costs

The results presented so far were obtained without considering real cost of items c^I. Let us introduce a more realistic cost by drawing c^I from the exponential distribution presented in figure 5.7. Let us set c^{OS} to 0.5 and c^{BO} to 5 for each period for each unit in back order. That respects the $\frac{c^{OS}}{c^{BO}} = 10\%$ used in (Babai et al., 2009) and (Syntetos et al., 2010b).

Again, even with the change in costs, the single selection does not change and SBA is still the selected model except for the SMRS based on sRMSE configuration where again, TSB is

selected. This is logic for the accuracy based selection since changing the cost of the items does not change the accuracy of the FM nor the decisions.

Table 5.5 Configurations' global performance with realistic costs

Selection	**model**	**CSL**	**SO** $\cdot 10^6$	**TC** $\cdot 10^6$	**mRank**
Single	SBA	94.7	2.555	129.4	13.500
	TSB	94.9	2.562	130.0	13.833
	arima	**98.0**	2.545	134.6	10.166
	snaive	97.9	2.495	134.2	8.000
	theta	96.4	2.650	137.3	13.000
MMPS	cost	**98.0**	2.556	120.8	9.500
	MASE	94.6	2.524	113.5	9.333
	sRMSE	94.9	2.534	113.9	9.500
	sAPIS	94.6	2.524	113.5	9.333
MMIS	cost	95.4	2.457	**109.5**	**3.833**
	MASE	95.3	**2.446**	112.2	4.333
	sRMSE	95.4	2.493	114.2	7.167
	sAPIS	94.6	2.544	114.4	12.667
MMIPS	cost	96.5	2.514	114.2	4.333
	MASE	96.3	2.475	113.8	7.333
	sRMSE	95.9	2.483	114.7	8.000
	sAPIS	94.7	2.530	114.0	11.500

Inspecting table 5.5, one can observe that SBA still outperforms the other single FM configurations in term of SO and TC.

One can also observe from figure 5.8 that when considering realistic costs, the impact of multi-selection appears even more important.

Considering again with realistic costs the lift induced by simulation (see table 5.6 for results), one would find that it seems to have between no impact and around 3% lift if the decider only takes into consideration CSL and SO which are the only real measures of impact. Otherwise, one can consider the classical accuracy selection have selected the FM of minimal cost.

For multi-model selection, considering realistic costs makes the average lift of selecting multiple models on TC go from 0.9% to 15.3%. Which is a significant savings for a company. Recomputing the average lift induced by multiple selection compared to single model selection, we obtain a 4.9% average lift on all global performance metrics with a maximum of 6.6% lift for the best configuration: MMIS.cost.

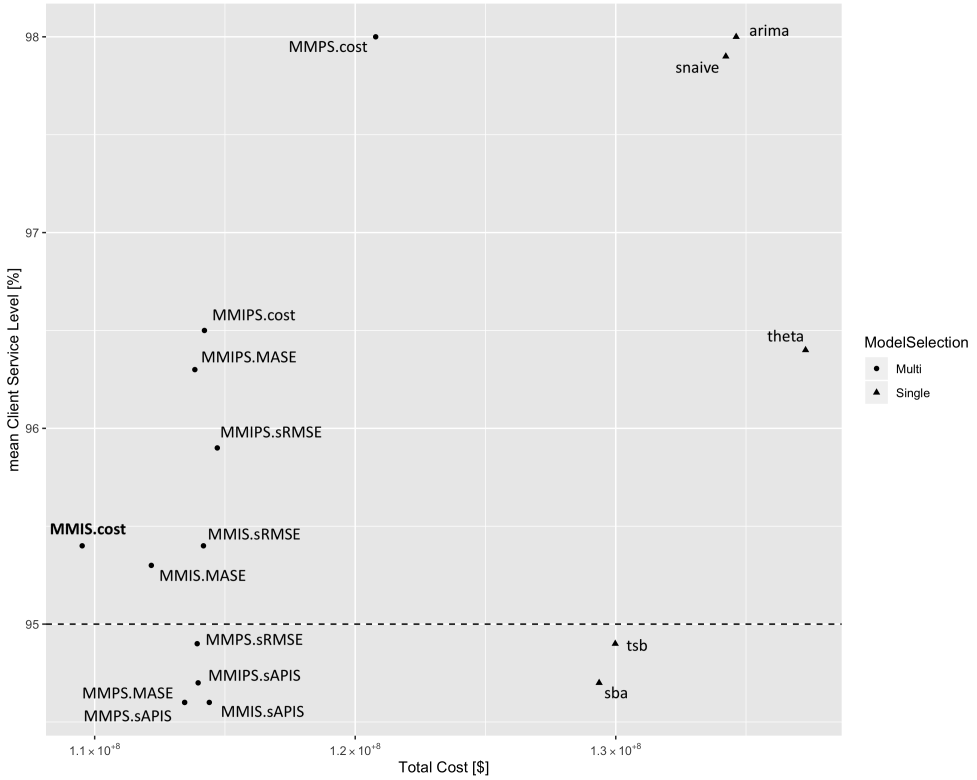

Figure 5.8 All Configurations TC compared to the average CSL with realistic costs

Table 5.6 Mean lift according to selection method with realistic costs

Selection	model	CSL (lift)	SO (lift)	TC (lift)	mean (lift)
Single	SBA	1	1	1	1
	TSB	0.998	1.003	1.005	1.002
	arima	**0.966**	0.996	1.041	1.001
	snaive	0.967	0.977	1.038	0.994
	theta	0.982	1.037	1.061	1.027
MMPS	cost	**0.966**	1.001	0.934	0.967
	MASE	1.001	0.988	0.877	0.955
	sRMSE	0.998	0.982	0.881	0.957
	sAPIS	1.001	0.988	0.877	0.955
MMIS	cost	0.993	0.962	**0.847**	**0.934**
	MASE	0.994	**0.958**	0.867	0.940
	sRMSE	0.993	0.976	0.883	0.951
	sAPIS	1.001	0.996	0.884	0.960
MMIPS	cost	0.981	0.984	0.883	0.949
	MASE	0.983	0.969	0.880	0.944
	sRMSE	0.987	0.972	0.887	0.949
	sAPIS	1.000	0.990	0.881	0.957

5.6 Conclusion

This paper proposed a methodology to select multiple forecasting models in an inventory management context. To do so, a simulation framework was proposed to measure inventory performance of different FM according to an IP. The framework allows for dynamic changes in both the FM and the IP for each item and periods. Using the proposed simulation framework, we have tested different selection methodologies. Among them, new multiple FM selections have been tested to estimate the impact in terms of lift of: 1. simulating the inventory performance of a single FM and 2. multiple selection. The results were validated by cross-validation on an independent set.

Based on service level, stock ordered, and total cost, the selection driven by simulation could not improve the results, as the most accurate FM was also the model of minimum cost. However, simulation could reveal other prospective models for selection, which required less stock (2% less) and exceeded the target CSL (4% increase). Otherwise, the impact of the simulation was minimal, since it selected the same model as classical accuracy metrics.

On the other hand, multi-model selection results showed a clear impact when allowing the FM selection to be made for each individual item (MMIS). Selecting the best FM at the item level increased the Service Level (7% increase), reduced the number of units ordered (4% reduction), and reduced the costs (15% reduction) if it was based on either the cost-based cost function or MASE. Indeed, the improvements were not as clear when selecting models based on sRMSE or sAPIS.

Based on the results, the cost of developing the simulator and the extra wait for the results might not bring significant improvements to the performances compared to classical or multiple model selection based on MASE. Nevertheless, the simulation remains useful to translate performance into cost and service units. To summarize, the results showed that a selection based on cost is more reliable than based on accuracy. Multiple selection also shown to improve results compared to single selection.

Since the goal of any organization should be to reduce their operations' costs while maintaining good service, it would be interesting to test if optimizing the IP given the FM can lead to greater impact on service and costs than an improvement of FM accuracy given the IP as we tested in this paper. That would be a step towards optimizing the FM and IP together.

Acknowledgments

The authors would like to acknowledge our industrial partner (Logistik Unicorp) and MITACS for funding this work under grant MITACS IT12058, and for providing other support.

CHAPITRE 6 ARTICLE 3: AN IMITATION LEARNING APPROACH TO INVENTORY MANAGEMENT

St-Aubin, P., Agard, B.

Abstract *- In many cases in inventory management, it may be possible to know what would have been the best inventory management decision given historical events. In this context, most previous works have focused on optimizing a predefined Inventory Policy (IP) to determine reorder decisions. Instead, in this paper, we propose to learn a dynamic reorder policy using Imitation Learning (IL), which is an approach used to convert an optimization problem into a supervised machine learning problem when optimal or near optimal solutions are known. The derived IL-based policy is compared to a static (s,Q) and a dynamic (s_t, Q) reorder policy. The policies were implemented and fed using four different forecasting models (seasonal arima, sba, snaive, theta). Our results showed that the most accurate Forecasting Models (FM) did not yield the best inventory performance and that using an IL-based policy could improve the inventory performance by almost 30% compared to the static (s,Q) policy.*

Keywords: Imitation learning, Inventory management, Inventory control, Machine learning, Dynamic inventory policy

6.1 Introduction

The performance of an inventory management system relies heavily on demand forecasts (Prak and Teunter, 2019). Therefore, changes in the Forecasting Model (FM) or in the performance of the FM can require that the Inventory Policy (IP) parameters be recalibrated to optimize inventory performance. Recalibrating the IP to the FM changes can be time consuming and complicated. One way to mitigate this is to use a policy that can adapt to changes in the performance of a FM. A technology that can add such information to adapt its prediction is Machine Learning (ML). Indeed, a ML model could be trained to learn the impact of changes in the performance of a FM and thus allow the IP to adapt its decision. However, the methodology to learn an IP using ML is not evident.

In this paper, we propose to use Imitation Learning (IL), an approach to transform control problems into supervised learning problems that can be solved with ML (Hussein et al., 2017). In the present case, IL is used to learn an IP depending on the state of the inventory and other factors. The methodology is tested with real data and the performance of the learned policy is compared to that of a static (s,Q) and a dynamic (s_t, Q) policy with its parameters

updated every period as this was shown to reduce inventory costs (Babai et al., 2009). The idea is to develop a dynamic IP that does not need to have its parameters adjusted whenever a change in the FM or in the performance of the FM occurs.

The following sections present actual approaches to solve the problem (section 6.2), the proposed methodology to approach an inventory management problem as an IL problem (section 6.3), the details of the conducted experiment (section 6.4) and our results (section 6.5).

6.2 Previous works

The problem of managing inventory has mainly been approached by selecting an IP and trying to optimize its parameters to minimize costs or stock levels while maintaining an acceptable service level. Different heuristics to select the parameters were proposed for (s,S) and (T,s,S) policies. Among the most known are Naddor's heuristic (Naddor, 1975), the Power approximation (Ehrhardt, 1979) and the Normal approximation (Wagner, 1970). However, no significant differences in performance were observed in an empirical comparison (Sani and Kingsman, 1997). Another empirical study Babai et al. (2010) found that the Normal approximation yielded the best results in terms of backlog/inventory holding. For the reorder point policy (r,Q), the Silver-Meal (Silver, 1973) or the EOQ (Wilson, 1934) heuristics can be used to calculate the order quantity Q and an hypothesis on the demand distribution can be made to determine the reorder point r (Axsäter, 2006). While these heuristics make approximations about the properties of the demand distribution, the costs, or the forecast error to derive expression of the parameters in function of the forecast error/demand variance (Syntetos et al., 2014) (Syntetos et al., 2010b), others such as Mohammaditabar et al. (2012) use meta heuristics or mixed integer linear programming (MILP) as Movahed and Zhang (2015) to determine when and how much to order.

So far, few studies present the impact of dynamic calculations of the parameters of an IP (do Rego and de Mesquita, 2015). Grewal et al. (2015) has shown through simulation that adapting the reorder point and quantity to seasonality could reduce inventory levels, and Babai and Dallery (2009) further showed that a dynamic reorder point (r,Q) policy reduced required inventory levels compared to a static policy. Even though extensive results about the impact of dynamic policies are not available, many authors believe it should reduce inventory levels. Among them, Kanet et al. (2010) advised for the use of dynamic parameters in the presence of varying service level requirements or non stationary demand or lead time. In most cases when a dynamic policy is proposed, it reevaluates the parameters at a certain frequency from one of the heuristic methods as in Tiacci and Saetta (2009), Babai et al.

(2009), and Syntetos et al. (2010b).

Another way to solve for dynamic policies is to approach the problem from an IL perspective. The idea of IL is to use demonstrations to train an agent (a learning machine) to perform a task (Hussein et al., 2017). Thus, if the optimal decision for a given situation (state) is known afterward from historical data, then one can collect a large set of (state, optimal decision) pairs to train the agent (a ML model). The model would map a state to its optimal decision resulting in a dynamic policy that adapts to the actual state. An advantage of IL is that using simulation to generate the states allows the agent to learn to cope with predictable changes of the environment or the task. In an inventory management context that means, the policy can learn the impact of a drift in performance of the FM, a change in the lead time, etc. Using IL also allows to include metadata such as the date or product group to learn a policy that considers covariates to improve performance. Things that are difficult to do with classical IP.

IL was used in Abbasi et al. (2020) who trained a ML model to reproduce optimal decisions obtained from solving a stochastic optimization problem on the logistics of sharing a blood supply among a network of hospitals. The ML model was able to reproduce near optimal solutions, reducing costs by around 29%. Similarly, Baniwal et al. (2019) could deduce from the structure of a problem related to predictive picking in a distribution center. IL was then used to reproduce the optimal policy. This approach performed much better than any other heuristics.

The following section presents an IL-based approach to model a dynamic IP with ML.

6.3 Methodology

This section presents the methodology to build and evaluate an IL-based IP. The general idea is to obtain state observations (Inventory level, Forecasts, etc.) noted $\phi(x_t)$ through simulation of a base heuristic (an (s,Q) IP). Then, given the actual demand is known, an optimal policy $\mu(x_t)$ can attribute the optimal decision u_t^* to each state observation $\phi(x_t)$. Once enough observations are gathered, a ML model is trained to learn a mapping (an approximation of the optimal policy noted $\tilde{\mu}$) of a state to a decision u.

All the components of the methodology are presented on figure 6.1. It presents the historical demand series (Y) of different items to manage including smooth series B and C and intermittent series A. Demand series (Y) are the only input data required to apply the methodology. The first step of the methodology is to partition historical demand in two subsets. The first, used to train the policy, includes all available periods t before t_v. Once a policy is learned,

the second subset, which includes all periods t greater or equal to t_v, is used to validate the generalization capacity of the policy on unseen data and to compare its performance to other benchmark policies.

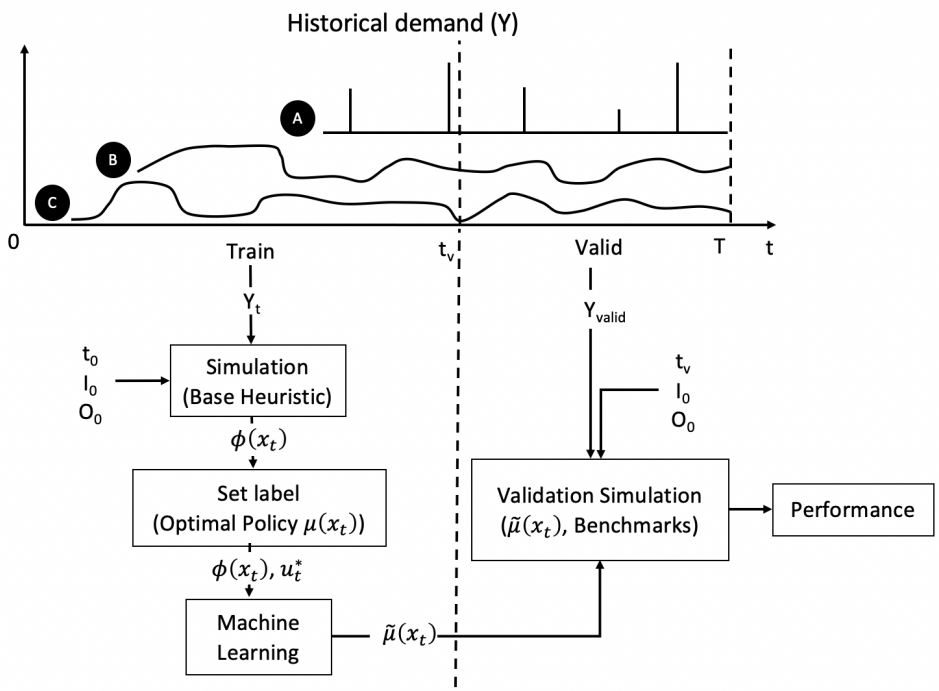

Figure 6.1 Imitation Learning and performance evaluation methodology

The following subsections divide the methodology in two. The first presents how to learn the policy applying IL on the training set. The second presents how to validate the IL-based policy using the validation set.

6.3.1 Learning the Inventory Policy

To train the policy, multiple labeled observations must be gathered. These observations can be obtained via simulation of a base heuristic. The following subsection describes the simulation and base heuristic used to gather observations.

6.3.1.1 Simulation

To gather multiple observations involving different states x_t of demand (Y_t), inventory (I_t), incoming orders (O_t), etc., one can use the historical demand and any IP as the base heuristic to drive the reorder process and simulate what would have happened given the actual demand

and the base heuristic. In this paper, an (s,Q) IP is used as base heuristic and the output of the simulation are observable features of the states $\phi(x_t)$. These are further detailed in the ML section 6.3.1.3. The simulation workflow is presented in figure 6.2 and detailed in this subsection.

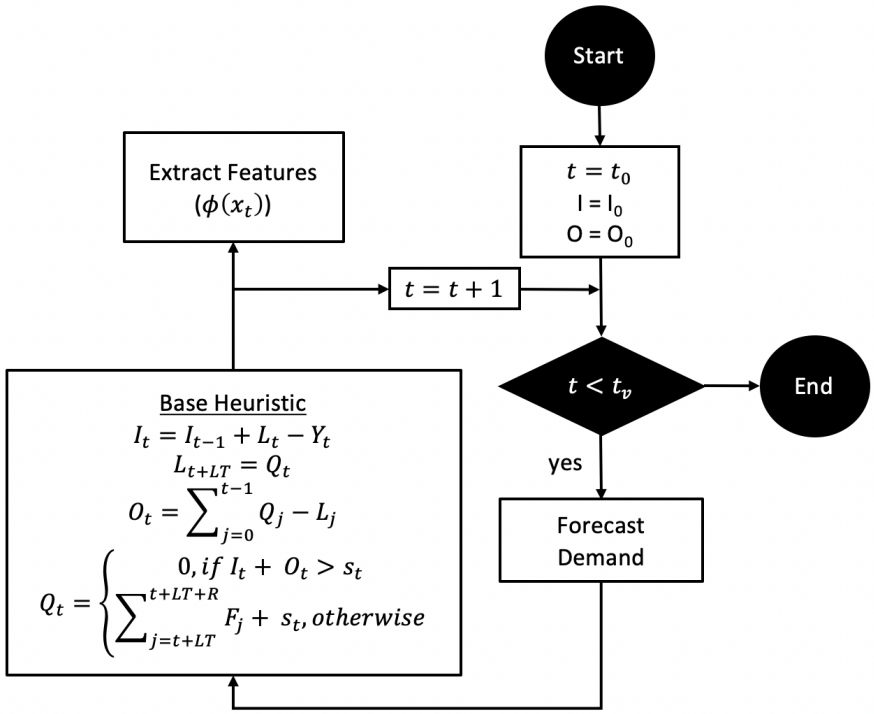

Figure 6.2 Simulation flow chart

First, the initial period t, the inventory I, and the incoming orders O are set equal to the initialization parameters (t_0, I_0, O_0). Since an (s,Q) policy is used, demand forecasts are required to determine the right quantity to order.

6.3.1.1.1 Forecast Demand For the simulation, four different FM are used: seasonal arima (arima), seasonal naive (snaive) (Hyndman and Athanasopoulos, 2018), theta (Assimakopoulos and Nikolopoulos, 2000), and SBA (Syntetos and Boylan, 2005). Their parameters are selected either to minimize the Akaike Information Criterion (AIC), if a statistical model is available, or to minimize the in-sample Mean Squared Error (MSE) otherwise. The FM's parameters are reevaluated every period and therefore use the data from period $t = 0$ to the actual period t to forecast demand over the required horizon. Following this, each FM's in-sample MSE at period t is evaluated (MSE_t) as this is used to calculate the parameters of the base heuristic.

6.3.1.1.2 Base Heuristic The base heuristic can be any classical IP. In this paper, the chosen IP is an (s,Q) policy. The reorder point is evaluated by:

$$s_t = \sum_{j=t}^{t+LT+R} F_j + \Phi^{-1}(TSL)\sqrt{MSE_t \cdot (LT+R)} \tag{6.1}$$

The first part of the equation represents the estimated demand the inventory should at all times be able to cover. F_j is the forecasted demand for period j. The second part of the equation is the safety stock (ss). The safety stock approximation follows the Normal approximation heuristic (Wagner, 1970) as it was shown to perform well in (Babai et al., 2010) and Sani and Kingsman (1997). Φ is the cumulative normal distribution; TSL is the Target Service Level (TSL); LT is the Lead Time (LT); and R is the number of periods an order should cover for.

The order quantity (Q) depends on the stock level, the incoming orders and the demand forecasts.

$$Q_t = \begin{cases} 0, \text{ if } I_t + O_t > s_t \\ \sum_{j=t+LT}^{t+LT+R} F_j + s_t - (I_t + O_t), \text{ otherwise.} \end{cases} \tag{6.2}$$

Where $O_t = \sum_{j=0}^{t-1} Q_j - L_j$ is the incoming orders (orders to be received); $L_{t+LT} = Q_t$ is the reception at period $t+LT$. The receptions are the lagged vector of the orders Q by LT. The inventory I_t is updated according to: $I_t = I_{t-1} + L_t - Y_t$.

After the base heuristic updates the inventory levels, some additional random noise can be added to the inventory levels to increase the probability of spanning all possible states.

6.3.1.1.3 Extract Features Once an iteration of the simulation is over, observable features $\phi(x_t)$ of the state x_t are saved as observations for the supervised learning problem to model the policy. The features used in this paper are the I_t, O_t, $F_t^s = \sum_{j=t}^{t+LT+R} F_j$, $F_t^Q = \sum_{j=t+LT}^{t+LT+R} F_j$, LT, t, $MSE_t(F)$, the latter being the in-sample MSE of the model trained at period t. These features are extracted for each iteration of t for each FM and each item. The observations are kept in memory to build the dataset to train the policy with ML. The next subsection discusses how the labels can be added to the dataset.

6.3.1.2 Set Label (optimal policy)

To approach an inventory management problem with IL, one needs to know what is the optimal decision in any given case. In this subsection we argue what such a policy could be to determine the labels for the dataset.

A company generally wishes to minimize its stock investment, modeled as:

$$J_{i,t} = u_{i,t} \cdot c^{I}_{i,t} \tag{6.3}$$

Where $u_{i,t}$ is the quantity ordered for the item i at period t, and $c^{I}_{i,t}$ is the unit cost of item i plus holding cost at period t. So, $J_{i,t}$ represents the cost at period t for the item i.

Another important factor to consider is the Service Level (SL) to balance out the objective to minimize stock. To take this into consideration, under and over-achieved TSL are attributed an additional penalty cost (P):

$$\begin{gathered} P_{i,t} = |I_{i,t} - TSL \cdot Y_{i,t}| \cdot \\ \left[\frac{c^{BO}_{i,t}}{2} \Big(1 - sgn(I_{i,t} - TSL \cdot Y_{i,t}) \Big) + \frac{c^{OS}_{i,t}}{2} \Big(1 + sgn(I_{i,t} - TSL \cdot Y_{i,t}) \Big) \right] \end{gathered} \tag{6.4}$$

And so the total cost is :

$$J_{i,t} = u_{i,t} \cdot c^{I}_{i,t} + P_{i,t} \tag{6.5}$$

In addition to the cost of the ordering decision $u_{i,t}$, the unit cost of under-achieving $c^{BO}_{i,t}$ and over-achieving $c^{OS}_{i,t}$ the TSL are added, where $I_{i,t}$ and $Y_{i,t}$ are the inventory level and the demand of item i at period t respectively. The function sgn is the Sign function. Considering the outcome of under-achieving the TSL can result in contract loss in addition to the usual extra costs of back ordering, $c^{BO}_{i,t}$ is greater than the other costs. Therefore, to minimize the function $J_{i,t}$, one needs to minimize the ordered stock to avoid over-achieving TSL costs and to minimize the ordering costs, while ordering enough stock to avoid the high extra cost of under-achieving the TSL.

In a context where suppliers allow multiple opened orders, an optimal policy would order, at every period, the stock required to fill TSL percent of the demand at $t + LT$ where LT is the Lead Time. Such a policy would render the TSL at a minimum cost. Some suppliers might not be willing to take new orders every period. In that case, an additional parameter R should be added to represent the number of periods the order must fill the demand for. As well, the orders should be passed every R periods to fill TSL percent of the demand between periods $t + LT$ and $t + LT + R$.

To summarize, the optimal reorder policy in a context where multiple orders are allowed, reorder the quantity required to fill the demand for R periods if it is no longer possible to wait to reorder before the Realized Service Level (RSL) becomes less than the TSL. That

can be written as the following rule :

$$u^*_{i,t} = \begin{cases} 0, \text{ if } I_{i,t} + O_{i,t} > TSL \cdot \sum_{j=t}^{t+LT+R} Y_{i,j} \\ TSL \cdot \sum_{j=t}^{t+LT+R} Y_{i,j} - (I_{i,t} + O_{i,t}), \text{ otherwise.} \end{cases} \tag{6.6}$$

Where $u^*_{i,t}$ represents the order quantity for item i at period t. An order must be passed if the inventory level with incoming orders cannot meet the demand for the following $LT + R$ periods. The quantity to reorder should be the difference between the demand between periods t and $t + LT + R$ and the inventory level with incoming orders.

This policy yields minimum costs if used with historical data where the actual demand $Y_{i,t}$ is known beforehand. Thus, this policy is used to label any given state of inventory, incoming orders, and demand for the following LT+R periods with the optimal decision.

6.3.1.3 Machine learning

From the labeled dataset, the extracted features of the state $\phi(x)$ (I_t, O_t, F^s_t, F^Q_t, LT, t, $MSE_t(F)$ in this case) and additional metadata are used to learn the reorder policy with supervised learning. Supervised learning uses ML to map a function (the policy $\tilde{\mu}$) of the input (the feature vector $\phi(x_t)$) to the output (the decision u^*_t) (Russell and Norvig, 2010).

Since the optimal policy is non linear, being 0 when the inventory and incoming stock is greater than the demand for the range to cover (see equation 6.6), the target variable is transformed into a linear equation to ease the learning of the policy. Another simplification to the optimal policy was performed to make the learned policy more versatile. The TSL term from the optimal policy was dropped in order to learn the exact stock required. This allows one to change the TSL at a later time and still have a valid policy. For this reason, the definition of u^*_t is adapted to include negative values and to incorporate a TSL of 1.

$$u^*_{i,t} = \sum_{j=t}^{t+LT+R} Y_{i,j} - (I_{i,t} + O_{i,t}) \tag{6.7}$$

Thus, the model learns the difference between the demand and the stock levels instead of learning the decision directly. To find the decision variable again, negative values of $\tilde{\mu}(x)$ are set to 0 to transform $\tilde{\mu}(x_t)$ into the decision u_t. The TSL can also be incorporated into the prediction afterward by taking:

$$u_t = TSL \cdot \tilde{\mu}(x_t) + (TSL - 1) \cdot (I_{i,t} + O_{i,t}) \tag{6.8}$$

6.3.2 Validation

Once a policy $\tilde{\mu}$ is learned, the validation subset of the historical demand including periods from t_v to the last period T is used to evaluate inventory performance.

6.3.2.1 Validation Simulation

The same simulation method is used to validate the performance of the IL based policy. This time however, the historical demand from the validation set (Y_{valid}) is used to feed the simulation with $\tilde{\mu}$ and the other benchmark methods as IP to drive the inventory ordering process. Once the simulation for the whole validation subset is done, the inventory performance of each policy is analyzed using different performance metrics.

6.3.2.2 Performance

To evaluate the performance of the different IP, three inventory based performance metrics are used:

Total Cost : $TC = \sum_{i=1}^{N} \sum_{j=t_v}^{T} J_{i,j}$. It sums the cost J over all periods j and items i.

Stock Ordered : $SO = \sum_{i=1}^{N} \sum_{j=t_v}^{T} Q_{i,j}$. It represents the total quantity of stock ordered for the whole validation period for all items.

Average Client Service Level : $\overline{CSL} = \frac{1}{N \cdot T} \sum_{i=1}^{N} \sum_{j=t_v}^{T} min(1, PI_{i,t}/Y_{i,t})$. It represents the RSL from the point of view of a client. Where the Physical Inventory $PI = max(0, I)$, it implies the CSL is between zero and one.

These metrics are used to compare and evaluate the performance of the different IP.

This concludes the presentation of the methodology. The following section presents an experiment in which the methodology is applied to a real dataset to evaluate its validity and compare its performance to classical IP.

6.4 Experiment and contribution

The experiment applies the methodology using the aforementioned simulation method and base heuristic. The main contribution of the experiment is to:

1. Confirm the validity of the methodology and the applicability of IL to inventory management.
2. Measure the lift induced by using dynamic policies versus static ones.
3. Compare the impact of improving the FM versus improving the IP.

This information will allow practitioners and researchers to know where to invest time and effort to maximize the impact on inventory performance, whether by putting effort into sophisticated dynamic policies or in the accuracy of Forecasting Models.

6.4.1 Data

The demand data for the experiment comes from a company that supplies uniforms for several different organizations. It contains the demand for over 10 000 items between 2012 and 2019. To conduct the experiment, the first two years of data are used to fit the forecasting models required for the simulation. So, t_0 corresponds to the first week of 2014. The last complete year of data is used to validate the methodology with t_v corresponding to the first week of 2018.

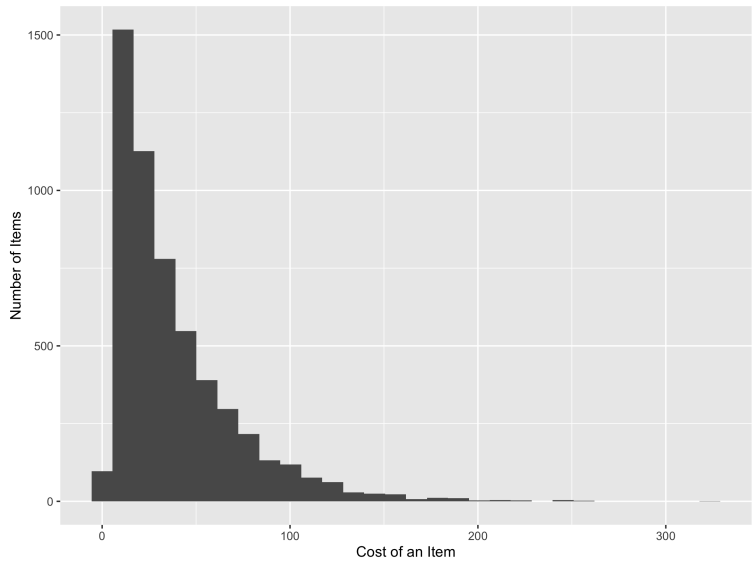

Figure 6.3 Items cost distribution

The costs used for the experiment are fictitious. The item costs (c^I) were drawn from an exponential distribution presented in figure 6.3. The penalty costs c^{OS} and c^{BO} were set to 0.5 and 5 respectively. This seems to be a reasonable estimation that respects the ratio $\frac{c^{OS}}{c^{BO}} = 10\%$ considered realistic according to Babai et al. (2009) and Syntetos et al. (2010b).

6.4.2 Simulation

The dataset was built running the simulation 10 times with random t_0, I_0, O_0. The base heuristic to generate the state was also switched randomly with a static one with values of s set randomly as a proportion of the ss: $s = \gamma * ss$ with $0 \leq \gamma \leq 6$.

6.4.3 Machine Learning

Additional metadata was added to the dataset to help learning the decision. The metadata consisted of the gender of the item, its product group, and its season, along with information about time: week and month number. The first three mentioned were one-hot encoded and the week and month number were fourier transformed to represent the cyclical nature of time, bringing the total number of features to 30 with around 3M observations.

Since the items have a great variance in scale, the features and the target variable were standardized by subtracting the mean and dividing by the standard deviation of the training set.

Three different ML models were trained using 90% of the observations, and an additional ensemble model stacking the three model was trained using the remaining 10%.

The first trained model was a linear regression model with lasso shrinkage (Tibshirani, 1996). To improve the simple linear regression model, additional squared and interaction terms were added as features. The regularization parameter was optimized using 5 folds Cross Validation (CV). The second model was a Gradient Boosting model with early stopping to prevent overfitting. The third model was a Multi Layer Perceptron (MLP). It was found to give the best results with three layers of 30, 15 and 1 neurons. The Adam optimizer with early stopping was used to optimize the network.

Once the three models reached their optimal performance, they were combined in an ensemble stacking of the predictions. The three models trained on 90% of the data were used to predict u on the remaining 10% from which a 20% of data was kept as test set for the ensemble model. The three predictions as well as the metadata were used to train a Random Forest and another Gradient Boosting model.

6.4.4 Benchmarks

Two different benchmark policies were used to compare the performance of the IL based policies. The first one was a static (s,Q) policy with $s = \mu + \Phi^{-1}(TSL)\sqrt{MSE \cdot (LT + R)}$ where μ is the mean demand. The mean demand (μ) was evaluated using all previous demand to t_v (first week of 2018) and MSE was evaluated using out-of-sample error from the previous year 2017 with the forecast models trained with the data previous to 2017. The second benchmark was the dynamic version of the (s,Q) policy as presented in section 6.3.1.1.2.

6.5 Results

The results present the performance as defined in section 6.3.2. But, since four different FM were used in the experiment, three different aggregations across the FM are used to characterize the performance of the IP. The performance of the FM is studied as well to conclude which model is the most accurate.

6.5.1 Forecasting Models Performance

To know which forecasting model would have been selected in a real life situation for the year 2018, the out-of-sample accuracy of the models is evaluated using the models fitted with the demand prior to 2017, and their performance is evaluated on the year 2017. Three different error metrics are used, and a ranking method is used to combine the three evaluations together.

The first one is the Mean Absolute Scaled Error ($MASE$) (Hyndman and Koehler, 2006)

$$MASE = \frac{1}{N}\sum_{i}^{N}\frac{\frac{1}{J}\sum_{j}|e_{i,j}|}{\frac{1}{T-m}\sum_{t=m+1}^{T}|Y_{i,t}-Y_{i,t-m}|} \tag{6.9}$$

Where $e_{i,j} = Y_{i,j} - F_{i,j}$ is the forecast error for item i at period j and the denominator is the in-sample MAE of the seasonal naïve forecast with a period of 1 year. The second metric is the scaled Root Mean Squared Error ($sRMSE$) where the scaling factor is given by the in-sample mean demand (Petropoulos and Kourentzes, 2015):

$$sRMSE = \frac{1}{N}\sum_{i}^{N}\frac{\sqrt{\frac{1}{J}\sum_{j}e_{i,j}^{2}}}{\frac{1}{T}\sum_{t}^{T}Y_{t}} \tag{6.10}$$

The third metric is the mean scaled Absolute Period In Stock ($msAPIS$) (Wallström and Segerstedt, 2010), (Kourentzes, 2014):

$$msAPIS = \frac{1}{N}\sum_{i}^{N}\frac{\sum_{h}^{H}\sum_{j}^{h}|e_{i,j}|}{\frac{1}{T}\sum_{t}^{T}Y_{t}} \tag{6.11}$$

Finally, the mean Rank ($mRank$) ranks the models relatively to each other for each metrics by series to forecast, then average the ranks over all metrics and series.

Table 6.1 Forecasting models accuracy (year 2017 out-of-sample)

model	**MASE**	**sRMSE**	**sAPIS**	**mRank**
arima	1.252	26.938	105.710	2.446
SBA	**1.003**	**17.262**	**78.156**	**1.612**
snaive	1.041	18.020	87.351	3.161
theta	1.033	17.653	86.705	2.781

Table 6.1 shows that SBA is the most accurate model for 2017 out-of-sample. Therefore, it would be advised to select SBA to feed the inventory management system.

To confirm the selection, performance is evaluated on the validation year as well. The model was trained with demand prior to 2018 and performance evaluated with 2018 out-of-sample.

Table 6.2 Forecasting models accuracy (year 2018 out-of-sample)

model	**MASE**	**sRMSE**	**sAPIS**	**mRank**
arima	1.333	32.156	60.713	2.433
SBA	**1.034**	**9.938**	**39.095**	**1.607**
snaive	1.088	11.076	39.368	3.192
theta	1.063	11.451	40.949	2.768

The results presented on table 6.2 shows that, indeed, SBA is the most accurate model for the validation period (year 2018), and the results seem stable across years.

6.5.2 Inventory Policies Performance

The inventory performance is first presented including the results of all combinations of IP and FM in table 6.3. The results show that the most accurate FM (SBA) systematically performs less well than the other FM with respect to the CSL. On the other metrics, SBA does not improve the performance except for the static policy. That shows the relation between inventory performance and forecasting accuracy is not direct and an increase in accuracy does not necessarily lead to improved inventory performance.

Since it was shown that a classical accuracy evaluation of the FM has concluded that SBA is the most accurate model, this model is kept for visualization. To take into consideration the robustness of the policy to changing performance of the FM, the mean and the standard deviation of performance according to the different FM are plotted in figure 6.4. The error bars represent a standard deviation on each side of the mean. The horizontal standard deviation is calculated according to the x axis variable and vertical to the y axis variable.

Table 6.3 Inventory Performance

policy	**model**	**CSL**	**SO** $\cdot 10^6$	**TC** $\cdot 10^6$	**mRank**
ensemble_GB	arima	0.991	1.586	77.459	6.33
	SBA	0.973	1.622	78.066	11.33
	snaive	0.990	1.536	78.461	6.33
	theta	0.986	1.671	84.168	13.50
ensemble_RF	arima	0.975	1.544	**74.834**	6.00
	SBA	0.930	1.600	76.772	14.67
	snaive	0.975	**1.499**	75.931	**4.67**
	theta	0.965	1.629	81.730	15.00
GB	arima	0.976	1.539	74.923	4.83
	SBA	0.932	1.593	76.268	13.33
	snaive	0.975	1.502	76.178	5.33
	theta	0.966	1.641	82.458	15.67
Lasso	arima	0.971	1.632	79.065	13.33
	SBA	0.946	1.709	81.722	19.33
	snaive	0.970	1.543	78.360	10.33
	theta	0.960	1.707	85.955	19.33
MLP	arima	0.952	1.571	76.569	12.00
	SBA	0.891	1.685	81.677	20.00
	snaive	0.943	1.516	77.594	12.00
	theta	0.921	1.654	83.868	20.33
s_t,Q	arima	0.986	2.124	108.004	17.17
	SBA	0.955	2.155	106.730	22.33
	snaive	0.976	2.035	106.870	16.83
	theta	0.962	2.220	113.275	23.68
s,Q	arima	**0.993**	2.116	111.239	16.33
	SBA	0.958	2.289	110.840	23.67
	snaive	0.980	2.091	112.233	18.00
	theta	0.964	2.423	122.910	24.33

The IL-based policy decreases the total level of stock ordered by around 26% to 30% compared to the static (s,Q) IP. While the dynamic (s_t,Q) reduces it by around 4%. However, figure 6.4 also show most IL-based policy reduce the mean CSL by around 1.5%. The mentioned lift is calculated from table 6.4, which was obtained by measuring the average performance over the FM and dividing it by those of the static policy.

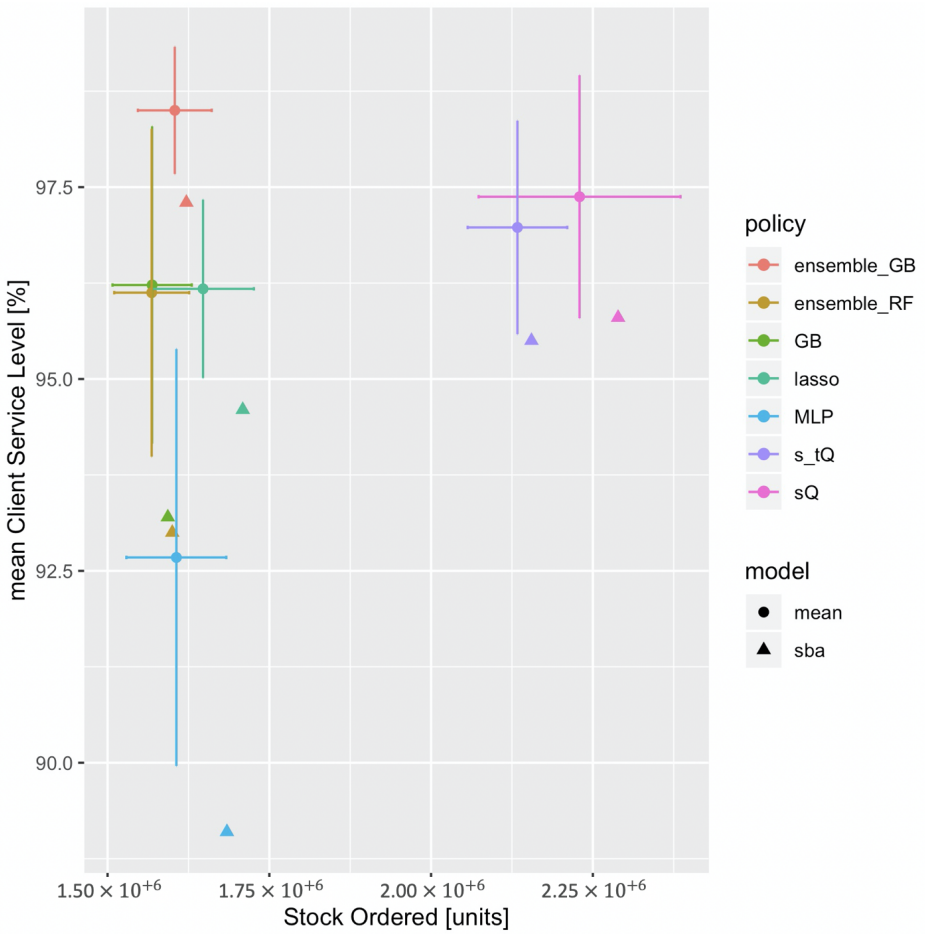

Figure 6.4 Policies mean and standard deviation across all FM and SBA Client Service Level and Stock Ordered

Table 6.4 Average Performance Lift

policy	CSL (lift)	SO (lift)	TC (lift)
ensemble_GB	**1.012**	0.719	0.696
ensemble_RF	0.987	**0.703**	**0.676**
GB	0.988	0.704	0.678
lasso	0.988	0.739	0.711
MLP	0.952	0.721	0.699
s_tQ	0.996	0.957	0.951
sQ	1.000	1.000	1.000

The same visualization is used to present the performance of the policies according to the total costs.

According to the inventory performance metrics, if a single combination policy x model were

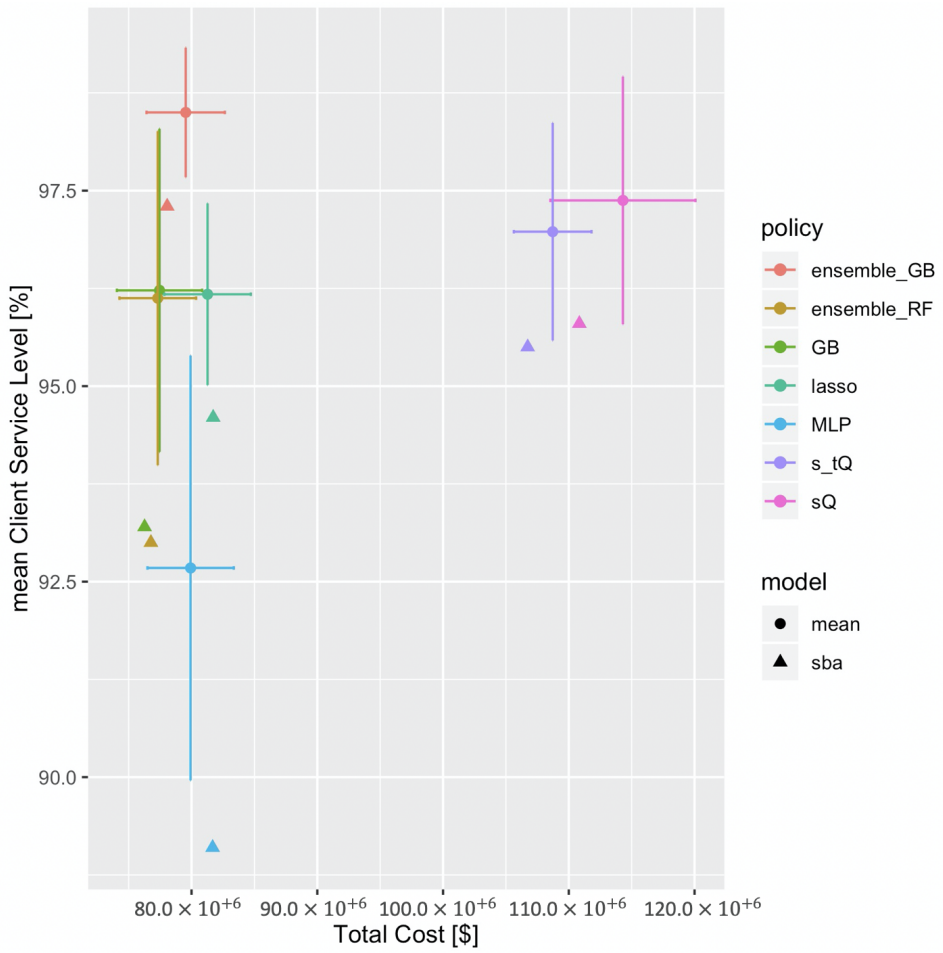

Figure 6.5 Policies mean and standard deviation across all FM and SBA Client Service Level and Total Cost

chosen, the combination with the lowest cost with RSL greater than TSL should be selected. However, considering a real life application, the FM's performance could change in time, or one might want to change the FM without having to simulate the whole inventory system. To take this in consideration, additional importance should be accorded to the stability of the performance of a policy. For this reason, from the results in table 6.3 and according to the visualization on figure 6.4 and 6.5, one can conclude the policy with the most stable results (smallest standard deviation) and best average results is the ensemble gradient boosting model.

6.6 Conclusion

In this paper, an IL-based methodology was proposed to learn a dynamic inventory policy. The idea of IL is to transform an optimization problem into a supervised learning problem.

To do so, a dataset with a large number of features of the sampled state with their optimal decision must be acquired. That was done by extracting observable features of states as the stock level, incoming orders, and other features. The states were obtained from a data driven simulation model from which the features were extracted at each iteration. The simulation model ran a base heuristic to drive the reorder process from the historical demand over 10k items to forecast demand for. The base heuristic was a dynamic (s_t, Q) policy similar to the one proposed in Babai et al. (2009). The forecasts were obtained from four different Forecasting Models (arima, sba, theta and snaive). Once a large enough dataset of sampled states was aquired, the optimal reorder decision was added as the target variable. This optimal decision could be retrieved given the structure of the costs and future demand was known. Once the optimal decision were obtained, ML could be used to train a policy to reproduce the optimal decisions in a given state.

That has led to the training of five IL-based policies a Lasso model, a Gradient Boosting model, a Multi Layer Perceptron, and two stacking ensemble models. The inventory performance of those policies were then compared to a classical static (s,Q) policy and to the base heuristic, the dynamic (s_t, Q) policy.

The results showed that an IL-based policy could reduce costs by around 30% compared to the static (s,Q) policy, while the (s_t, Q) policy could deliver only a 5% cost reduction. However, the IL-based policies also reduced the Realized Service Level by 1.2% on average. Yet, the ensemble Gradient Boosting model could keep stable results regardless of the FM used. To summarize, this paper showed that systematically updating IP parameters had a beneficial impact over static parameters.

Another important finding from this paper is that improving the accuracy of the FM did not necessarily lead to improved inventory performance. Therefore, it appears that putting effort into optimizing the reorder policy yields greater improvements for inventory performance than working to improve the accuracy of the forecasts. For this reason, additional effort should go into the selection of FM in an inventory management context to guide practitioners and help researchers understand the relation between forecasts and inventory performance.

Acknowledgments

The authors would like to acknowledge our industrial partner (Logistik Unicorp) and MITACS for funding this work under grant MITACS IT12058, and for providing other support for this research.

CHAPITRE 7 DISCUSSION GÉNÉRALE

Les trois chapitres précédents ont présenté les résultats de nos recherches pour la résolution de certaines problématiques en lien avec la conception d'un système de gestion de l'inventaire. Ce chapitre met en perspectives les résultats obtenus et les présentent en relation les uns avec les autres.

D'abord, une méthode pour évaluer les métriques de performances de modèle de prévisions à été proposée. Des recommandations sur les métriques de performance à utiliser selon les circonstances ont également été données. Auparavant on ne pouvait qu'argumenter sur les propriétés théoriques des métriques ou faire des comparaisons entre les résultats avec l'hypothèse que les résultats partagés par le plus de métriques sont les "meilleurs". Cette contribution se distingue donc des travaux précédents puisqu'elle implique l'utilisation de résultats connus pour tirer des conclusions sur les propriétés des métriques.

Ensuite, une méthode de sélection de modèles en contexte de gestion de l'inventaire a permis d'apporter des réponses sur l'impact de l'agrégation dans la sélection. Ce qui est, à notre connaissance, la première étude menée sur des séries à profil mixte et la première à en évaluer les impacts en lien à la gestion d'inventaire. Notre recherche introduit également une terminologie claire pour classifier les méthodes de sélection. Cinq méthodes de sélection ont été identifiées.

Finalement, une méthode pour optimiser une politique d'inventaire dynamique adaptative aux performances d'un modèle de prévision, basée sur le cadre conceptuel de l'apprentissage par imitation, a été évaluée. Les résultats ont montré que l'approche proposée fonctionne et permet d'améliorer les résultats par rapport à une politique (r_t, Q). Les résultats semblent également montrer qu'il est plus avantageux de travailler à améliorer les politiques d'approvisionnement que les méthodes de prévisions puisque la variance sur les résultats des différents modèles de prévision est plus faible que celle sur les différentes politiques d'inventaire.

Par ailleurs, les résultats ont apporté de l'information supplémentaire sur la relation entre l'erreur de prévision et les performances d'inventaire. Il apparait que les relations trouvées précédemment pour des profils de demande lisse ne tiennent pas en présence de profils mixtes et que l'absence de corrélation identifiée dans des travaux précédents sur la demande intermittente se généralise au cas mixte.

Ce constat, avec le fait qu'il ait été identifié d'abord sur la demande intermittente spécifiquement, nous permet d'identifier pourquoi la relation entre la précision et les performances

d'inventaire est complexe à partir d'un exemple simple présenté sur la figure 7.1.

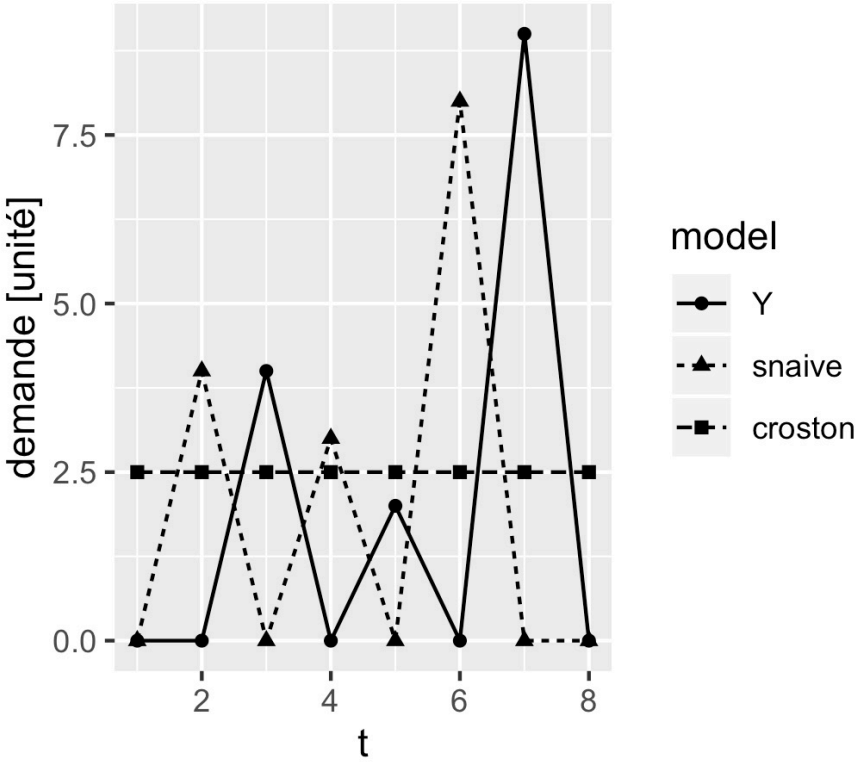

Figure 7.1 Exemple de prévision pour une série intermittente

La figure 7.1 présente des prévisions pour une série intermittente fictive Y. Soit un modèle de prévision qui essaie de reproduire les pics de demande de Y comme *snaive* et un modèle intermittent, ici, *croston*. Si on mesure la précision avec des métriques de performance basées sur l'erreur, on obtient les résultats présentés à la table 7.1.

Tableau 7.1 Précisions des modèles de prévision

model	**MAE**	**RMSE**	**APIS**
snaive	3.750	4.873	14.625
croston	2.625	3.082	10.500

Les résultats de la table 7.1 indiquent que la méthode de Croston est plus précise de 47% si on considère le lift moyen sur les trois métriques. Ce qui devrait garantir une bien meilleure performance en inventaire que la méthode snaive. Toutefois, pour obtenir une bonne performance en inventaire, un stock suffisant doit être disponible pour atteindre le niveau de service cible tout en ayant le minimum de stock nécessaire. Pour représenter cela, on considère une

politique qui commande le stock prévu à chaque période sans délai de livraison. On trace sur la figure 7.2 le stock à chaque période.

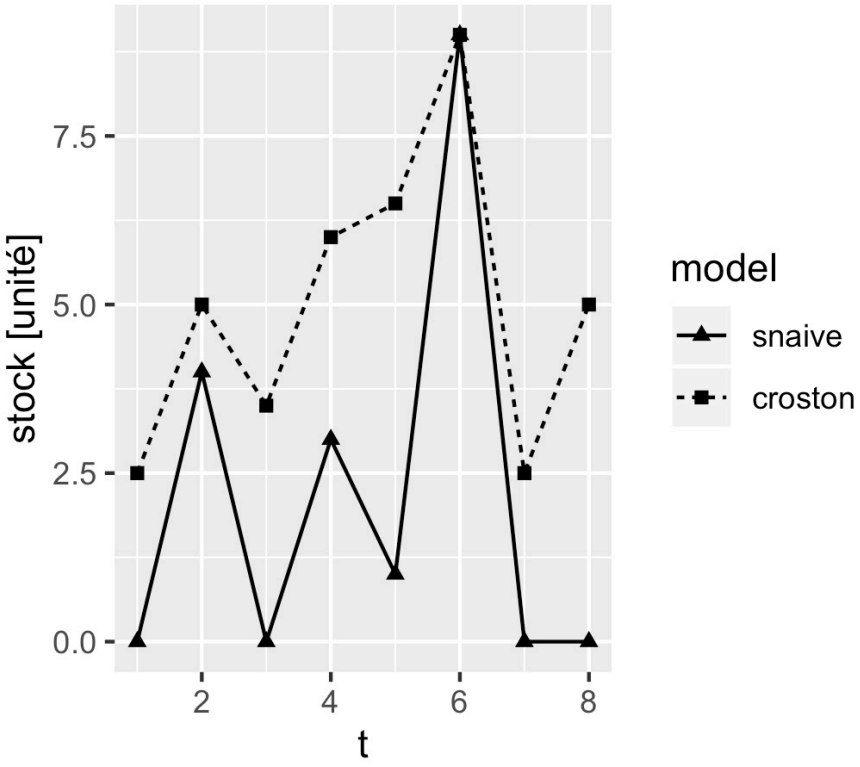

Figure 7.2 Quantité en stock à chaque période selon les recommandations du modèle

La figure 7.2 montre que la méthode de Croston implique de conserver un stock plus élevé en tout temps que la méthode snaive. Si on considère un coût de stockage (cos) et un coût de rupture de stock (cbo) par unité et par période symétrique ($cos = 1$, $cbo = 1$), alors on obtient un coût de 17$ pour snaive et de 40$ pour Croston. Ce qui confirme que la précision ne garantit pas une bonne performance en inventaire puisque ce qui compte réellement pour la gestion de l'inventaire est d'avoir le moins de stock possible pour pouvoir atteindre le niveau de service cible.

Pour considérer cela, une métrique basée sur l'erreur devrait pouvoir considérer le décalage entre les prévisions et la demande. Une piste de solution à ce problème est la métrique PIS qui permet de suivre un stock fictif. Toutefois, cette métrique ne donne que le niveau de stock final et est aveugle aux performances intermédiaires. Plus de recherche est donc requise pour développer une métrique de performance adéquate pour la gestion d'inventaire qui ne requiert pas de simulation.

Donc en plus de répondre aux questions de recherche, la mise en commun des méthodes

proposées avec les systèmes développés décrits au chapitre 3 permet de concevoir un système de gestion de l'inventaire efficace.

C'est surement en partie ce facteur qui permet d'expliquer pourquoi la méthode de Croston n'a pas permis d'améliorer les performances en pratique (Syntetos and Boylan, 2001), (Sani and Kingsman, 1997).

CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS

L'accessibilité grandissante pour un large public à de nouvelles méthodes de prévision toujours plus performantes suggère que la prochaine génération de systèmes de gestion de l'inventaire reposera en grande partie sur leur utilisation.

Toutefois, des recommandations claires sur plusieurs aspects de l'intégration des méthodes de prévisions aux systèmes de gestion de l'inventaire manquent toujours. Tel qu'identifié en revue de littérature, des recommandations manquent spécifiquement en présence de séries de demande à profil mixte (lisse et intermittent). Cette absence de recommandations combinée au contexte de démocratisation d'outils de prévision a mené à l'objectif générale visant la conception et le développement d'un système de gestion de l'inventaire pour un portefeuille de produits à profil de demande mixte. Plus spécifiquement, elle a inspiré trois objectifs spécifiques en lien avec le contexte d'application et le besoin de recommandations en présence de séries à profil mixte. Ces trois objectifs spécifiques mis bout à bout forment un cadre conceptuel pour guider l'utilisation de méthodes de prévisions dans la conception d'un système de gestion de l'inventaire.

Les contributions liées à la résolution de chaque objectif spécifique puis ceux de la thèse dans son ensemble sont présentées dans cette section.

Le premier objectif spécifique présenté au chapitre 4 vise à développer une méthodologie d'évaluation des métriques de performance pour les modèles de prévision. Pour ce faire, une méthodologie se basant sur l'utilisation d'erreurs connues a été proposée. Ce faisant, l'étude et l'analyse des propriétés des métriques de performance pouvaient maintenant se baser sur des indicateurs clairs. Une méthode pour évaluer la sensibilité et la fiabilité des métriques a ensuite été testée dans différentes circonstances, afin de comparer les propriétés de métriques de performance populaires pour évaluer les modèles de prévision.

Une critique de cette approche est qu'en pratique les erreurs d'un modèle n'ont pas la même distribution pour toutes les séries. Il aurait donc été intéressant d'étudier les propriétés des métriques pour une seule série à la fois. De même, la distribution de l'erreur est généralement croissante en fonction de l'horizon de prévision. Cet aspect a été négligé dans nos recherches.

Le second objectif spécifique présenté au chapitre 5 veut identifier une méthode de sélection des modèles de prévision dans un contexte de gestion de l'inventaire pour des items à profil mixte. Pour ce faire, différents scénarios de sélection ont été proposés et testés par simulation. L'impact de l'agrégation dans la sélection de modèles ainsi que la relation entre les métriques

de précision et les performances d'inventaire ont pu être mieux compris grâce à de nouveaux résultats. Les scénarios de sélection proposés impliquent dans le cas de la sélection multiple des paramètres. Ces paramètres n'ont pas été sélectionnés de manière optimale dans le cadre de la recherche. De meilleures performances pourraient être trouvées en sélectionnant les paramètres optimaux des méthodes de sélection multiple.

Finalement, le troisième objectif vise à tirer profit des nouvelles méthodes de prévision pour concevoir une politique d'inventaire dynamique en se basant sur le cadre conceptuel de l'apprentissage par imitation. Présentée au chapitre 6, cette méthode lance plusieurs simulations avec différents paramètres d'une politique d'inventaire de base et différents modèles de prévision de la demande pour acquérir plusieurs observations de différents états de l'inventaire. À chaque observation de l'état, une décision optimale est assignée pour former des paires observations - décision optimale. Ces paires ont ensuite été utilisées pour entraîner un modèle d'apprentissage automatique et ainsi apprendre une politique de réapprovisionnement dynamique qui s'adapte aux performances changeantes d'un modèle de prévision. Il aurait été intéressant de comparer cette approche pour déterminer une politique d'approvisionnement à une méthode d'apprentissage par renforcement comme un processus de décision markovien.

Pour les deux derniers objectifs, un délai de livraison stochastique n'a pas été considéré. Différentes méthodes de calculs du stock de sécurité auraient pu être utilisées. L'impact du niveau de service cible aurait aussi pu être étudié. Ajouter l'impact de ces facteurs pourrait permettre de conclure avec plus de certitude que les résultats observés sont dus à la variation des facteurs à l'étude. Une autre limite est l'absence d'information réelle sur les coûts et le choix des métriques d'inventaire qui ne permettent pas de connaître le niveau de stock moyen.

Chacune des propositions a été implémentée à partir de données provenant d'un cas d'étude industriel réel. D'ailleurs, les développements et les connaissances acquises par nos travaux de recherche sont en cours d'implémentation chez notre partenaire industriel. Donc, tous ensemble, les résultats, méthodes et recommandations faites dans cette thèse permettent de guider un ingénieur pour le développement d'un système de gestion de l'inventaire.

Les cas industriels sont la plupart du temps beaucoup plus complexes que la manière dont ils sont représentés dans les cas d'études en recherche. Notamment, il existe souvent plusieurs fournisseurs pour un même produit avec des prix différents et des délais de livraison différents. Dans nos études de cas, ces facteurs ont été ignorés et les données des fournisseurs à prioriser par le partenaire industriel ont été sélectionnées.

Une autre limite sur l'approche globale est que l'impact de l'ensemble des recommandations sur les performances d'inventaire n'a pas été étudié conjointement. Donc l'effet cumulé de la

sélection multiple avec l'utilisation d'une politique basée sur le cadre conceptuel de l'apprentissage par imitation reste inconnu.

L'analyse des résultats mis en commun a également pu révéler la raison faisant en sorte que des résultats mitigés étaient observés au sujet de la relation entre les performances de gestion d'inventaire et la précision. Ceci ouvre la porte à plusieurs opportunités de recherche sur le développement d'une métrique de performance capable de considérer les facteurs importants énoncés au chapitre 7.

Développer une telle métrique serait également une opportunité intéressante d'application de la méthodologie développée au chapitre 4. Avec cette méthodologie, la fiabilité des métriques de performance pour la sélection de modèles de prévision performants pour la gestion d'inventaire pourrait donc être quantifiée.

Une autre perspective de recherche intéressante serait d'intégrer le délai de livraison stochastique dans les méthodologies des chapitres 5 et 6 pour généraliser les conclusions à plus de cas rencontrés en industrie.

D'un point de vue global, certaines étapes de la méthodologie proposée pour la conception d'un système de gestion de l'inventaire autonome demeurent inexplorées. Ces étapes constituent des opportunités de recherche qui sont encore peu explorées jusqu'à maintenant et dont les impacts sont importants.

Ces étapes inexplorées de la méthodologie permettraient de garantir que les prévisions de la demande et les politiques d'inventaire seraient capables de s'adapter à une perturbation majeure dans les distributions de la demande. Cela rendrait le système robuste à des pannes dans le système ou encore à des évènements plus rares comme la situation actuelle où la pandémie mondiale de COVID-19 a modifié drastiquement la demande dans plusieurs secteurs de l'économie.

Les limites industrielles suggèrent que la situation des entreprises est souvent beaucoup plus complexe que celle modélisée dans les études de cas et sur lesquels reposent plusieurs résultats en gestion de l'inventaire. Cela permet de relever une question qui sera de plus en plus importante à poser dans les années à venir. Dans un contexte où le développement d'outils performants de contrôle optimal, d'apprentissage par renforcement, de prévision et d'apprentissage automatique est de mieux en mieux compris et où les outils requis pour développer des systèmes qui reposent sur ces outils se démocratisent, serait-il avantageux pour les entreprises de s'adapter pour faciliter l'utilisation et l'implémentation de ces technologies ? Où devrait-on adapter l'utilisation des technologies aux situations des entreprises ?

RÉFÉRENCES

B. Abbasi, T. Babaei, Z. Hosseinifard, K. Smith-Miles, et M. Dehghani, “Predicting solutions of large-scale optimization problems via machine learning : A case study in blood supply chain management”, *Computers & Operations Research*, p. 104941, 2020.

J. S. Armstrong et F. Collopy, “Error measures for generalizing about forecasting methods : Empirical comparisons”, *International Journal of Forecasting*, vol. 8, no. 1, pp. 69–80, 1992.

J. Arnold, *Introduction to materials management*. Pearson Prentice Hall, 2008.

V. Assimakopoulos et K. Nikolopoulos, “The theta model : a decomposition approach to forecasting”, *International Journal of Forecasting*, vol. 16, no. 4, pp. 521–530, 2000.

S. Axsäter, *Inventory control*. Springer, 2006.

M. Z. Babai, A. A. Syntetos, et R. Teunter, “On the empirical performance of (t, s, s) heuristics”, *European Journal of Operational Research*, vol. 202, no. 2, pp. 466–472, 2010.

M. Z. Babai, M. M. Ali, et K. Nikolopoulos, “Impact of temporal aggregation on stock control performance of intermittent demand estimators : Empirical analysis”, *Omega*, vol. 40, no. 6, pp. 713–721, 2012.

M. Z. Babai et Y. Dallery, “Dynamic versus static control policies in single stage production-inventory systems”, *International Journal of Production Research*, vol. 47, no. 2, pp. 415–433, 2009.

M. Z. Babai, A. A. Syntetos, Y. Dallery, et K. Nikolopoulos, “Dynamic re-order point inventory control with lead-time uncertainty : analysis and empirical investigation”, *International Journal of Production Research*, vol. 47, no. 9, pp. 2461–2483, 2009.

M. Babaï et Y. Dallery, “Inventory management : forecast based approach vs. standard approach”, dans *Proceedings of International Conference on Industrial Engineering and Systems Management, Marrakech*, 2005, pp. 1–10.

E. Babiloni, M. Cardos, J. M. Albarracin, et M. E. Palmer, “Demand categorisation, forecasting, and inventory control for intermittent demand items”, *South African Journal of Industrial Engineering*, vol. 21, no. 2, pp. 115–130, 2010.

V. Baniwal, C. Kayal, D. Shah, P. Ma, et H. Khadilkar, “An imitation learning approach for computing anticipatory picking decisions in retail distribution centres”, dans *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 4186–4191.

D. K. Barrow et N. Kourentzes, “Distributions of forecasting errors of forecast combinations : implications for inventory management”, *International Journal of Production Economics*, vol. 177, pp. 24–33, 2016.

B. Billah, M. L. King, R. D. Snyder, et A. B. Koehler, “Exponential smoothing model selection for forecasting”, *International Journal of Forecasting*, vol. 22, no. 2, pp. 239–247, 2006.

G. E. Box et G. M. Jenkins, “Time series analysis : forecasting and control holden-day san francisco”, *BoxTime Series Analysis : Forecasting and Control Holden Day*, 1970.

R. G. Brown, *Statistical forecasting for inventory control*. McGraw/Hill, 1959.

J. Bruzda, “Demand forecasting under fill rate constraints—the case of re-order points”, *International Journal of Forecasting*, 2020.

R. Carbone et J. S. Armstrong, “Note. evaluation of extrapolative forecasting methods : results of a survey of academicians and practitioners”, *Journal of Forecasting*, vol. 1, no. 2, pp. 215–217, 1982.

S. Çetinkaya et M. Parlar, “Nonlinear programming analysis to estimate implicit inventory backorder costs”, *Journal of Optimization Theory and Applications*, vol. 97, no. 1, pp. 71–92, 1998.

H. Chen, M. Z. Frank, et O. Q. Wu, “What actually happened to the inventories of american companies between 1981 and 2000 ?” *Management Science*, vol. 51, no. 7, pp. 1015–1031, 2005.

A. J. Clark et H. Scarf, “Optimal policies for a multi-echelon inventory problem”, *Management Science*, vol. 6, no. 4, pp. 475–490, 1960.

L. C. Coelho et G. Laporte, “Optimal joint replenishment, delivery and inventory management policies for perishable products”, *Computers & Operations Research*, vol. 47, pp. 42–52, 2014.

J. D. Croston, “Forecasting and stock control for intermittent demands”, *Journal of the Operational Research Society*, vol. 23, no. 3, pp. 289–303, 1972.

J. R. do Rego et M. A. de Mesquita, “Demand forecasting and inventory control : A simulation study on automotive spare parts”, *International Journal of Production Economics*, vol. 161, pp. 1–16, 2015.

R. Ehrhardt, “The power approximation for computing (s, s) inventory policies”, *Management Science*, vol. 25, no. 8, pp. 777–786, 1979.

E. S. Gardner, “Evaluating forecast performance in an inventory control system”, *Management Science*, vol. 36, no. 4, pp. 490–499, 1990.

B. Ghalebsaz-Jeddi, B. C. Shultes, et R. Haji, “A multi-product continuous review inventory system with stochastic demand, backorders, and a budget constraint”, *European Journal of Operational Research*, vol. 158, no. 2, pp. 456–469, 2004.

I. Giannoccaro et P. Pontrandolfo, “Inventory management in supply chains : a reinforcement learning approach”, *International Journal of Production Economics*, vol. 78, no. 2, pp. 153–161, 2002.

P. Goodwin et R. Lawton, “On the asymmetry of the symmetric mape”, *International Journal of Forecasting*, vol. 15, no. 4, pp. 405–408, 1999.

C. S. Grewal, P. Rogers, et S. Enns, “Performance evaluation of inventory replenishment strategies in a capacitated supply chain under optimal parameter settings”, *International Journal of Value Chain Management*, vol. 4, no. 3, pp. 195–212, 2010.

C. S. Grewal, S. T. Enns, et P. Rogers, “Dynamic reorder point replenishment strategies for a capacitated supply chain with seasonal demand”, *Computers & Industrial Engineering*, vol. 80, pp. 97–110, 2015.

F. W. Harris, “How many parts to make at once”, *Factory, The Magazine of Management*, vol. 10, no. 2, pp. 135–136, 1913.

C. Holt, “Forecasting trends and season als by exponentially weighted averages”, *Carnegie Institute of Technology, Pittsburgh*, 1957.

J. Hoover *et al.*, “How to track forecast accuracy to guide forecast process improvement”, *Foresight*, vol. 14, pp. 17–23, 2009.

A. Hussein, M. M. Gaber, E. Elyan, et C. Jayne, “Imitation learning : A survey of learning methods”, *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

R. J. Hyndman, “Measuring forecast accuracy”, *Business Forecasting : Practical Problems and Solutions*, pp. 177–183, 2014.

R. J. Hyndman et G. Athanasopoulos, *Forecasting : principles and practice.* OTexts, 2018.

R. J. Hyndman et B. Billah, “Unmasking the theta method”, *International Journal of Forecasting*, vol. 19, no. 2, pp. 287–290, 2003.

R. J. Hyndman et A. B. Koehler, “Another look at measures of forecast accuracy”, *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

R. J. Hyndman, A. B. Koehler, R. D. Snyder, et S. Grose, “A state space framework for automatic forecasting using exponential smoothing methods”, *International Journal of Forecasting*, vol. 18, no. 3, pp. 439–454, 2002.

K. Järvelin et J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques”, *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

J. J. Kanet, M. F. Gorman, et M. Stößlein, “Dynamic planned safety stocks in supply networks”, *International Journal of Production Research*, vol. 48, no. 22, pp. 6859–6880, 2010.

A. Kara et I. Dogan, “Reinforcement learning approaches for specifying ordering policies of perishable inventory systems”, *Expert Systems with Applications*, vol. 91, pp. 150–158, 2018.

S. Karlin, “Dynamic inventory policy with varying stochastic demands”, *Management Science*, vol. 6, no. 3, pp. 231–258, 1960.

C. Kim, J. Jun, J. Baek, R. Smith, et Y.-D. Kim, “Adaptive inventory control models for supply chain management”, *The International Journal of Advanced Manufacturing Technology*, vol. 26, no. 9-10, pp. 1184–1192, 2005.

M. Kırcı, I. Biçer, et R. W. Seifert, “Optimal replenishment cycle for perishable items facing demand uncertainty in a two-echelon inventory system”, *International Journal of Production Research*, vol. 57, no. 4, pp. 1250–1264, 2019.

A. V. Kostenko et R. J. Hyndman, “A note on the categorization of demand patterns”, *Journal of the Operational Research Society*, vol. 57, no. 10, pp. 1256–1257, 2006.

N. Kourentzes, “Intermittent demand forecasts with neural networks”, *International Journal of Production Economics*, vol. 143, no. 1, pp. 198–206, 2013.

——, “On intermittent demand model optimisation and selection”, *International Journal of Production Economics*, vol. 156, pp. 180–190, 2014.

N. Kourentzes, J. R. Trapero, et D. K. Barrow, “Optimising forecasting models for inventory planning”, *International Journal of Production Economics*, p. 107597, 2019.

S. Makridakis et M. Hibon, “The m3-competition : results, conclusions and implications”, *International Journal of Forecasting*, vol. 16, no. 4, pp. 451–476, 2000.

S. Makridakis, E. Spiliotis, et V. Assimakopoulos, “The m4 competition : Results, findings, conclusion and way forward”, *International Journal of Forecasting*, vol. 34, no. 4, pp. 802–808, 2018.

D. Mohammaditabar, S. H. Ghodsypour, et C. O’Brien, “Inventory control system design by integrating inventory classification and policy selection”, *International Journal of Production Economics*, vol. 140, no. 2, pp. 655–659, 2012.

K. K. Movahed et Z.-H. Zhang, “Robust design of (s, s) inventory policy parameters in supply chains with demand and lead time uncertainties”, *International Journal of Systems Science*, vol. 46, no. 12, pp. 2258–2268, 2015.

E. Naddor, “Optimal and heuristic decisions in single-and multi-item inventory systems”, *Management Science*, vol. 21, no. 11, pp. 1234–1249, 1975.

F. Petropoulos et N. Kourentzes, “Forecast combinations for intermittent demand”, *Journal of the Operational Research Society*, vol. 66, no. 6, pp. 914–924, 2015.

D. Prak et R. Teunter, “A general method for addressing forecasting uncertainty in inventory models”, *International Journal of Forecasting*, vol. 35, no. 1, pp. 224–238, 2019.

S. J. Russell et P. Norvig, *Artificial Intelligence : A Modern Approach.* Pearson Education London, 2010.

N. R. Sanders et G. A. Graman, “Quantifying costs of forecast errors : A case study of the warehouse environment”, *Omega*, vol. 37, no. 1, pp. 116–125, 2009.

B. Sani et B. G. Kingsman, “Selecting the best periodic inventory control and demand forecasting methods for low demand items”, *Journal of the Operational Research Society*,

vol. 48, no. 7, pp. 700–713, 1997.

E. A. Silver, “A heuristic for selecting lot size quantities for the case of a deterministic time-varying demand rate and discrete opportunities for replenishment”, *Production and Inventory Management*, vol. 2, pp. 64–74, 1973.

E. A. Silver, D. F. Pyke, R. Peterson *et al.*, *Inventory management and production planning and scheduling*. Wiley New York, 1998, vol. 3.

A. O. Solis, “Better statistical forecast accuracy does not always lead to better inventory control efficiency : the case of lumpy demand”, dans *14th International Conference on Modeling and Applied Simulation (MAS 2015)*, A. G. Bruzzone, F. De Felice, C. Frydman, M. Massei, Y. Merkuryev, et A. Solis, éds., September 2015, pp. 211–217.

C. Spearman, “The proof and measurement of association between two things”, *The American Journal of Psychology*, vol. 100, no. 3/4, pp. 441–471, 1987.

L. W. Strijbosch, A. A. Syntetos, J. E. Boylan, et E. Janssen, “On the interaction between forecasting and stock control : The case of non-stationary demand”, *International Journal of Production Economics*, vol. 133, no. 1, pp. 470–480, 2011.

A. Syntetos, M. Babai, J. Davies, et D. Stephenson, “Forecasting and stock control : A study in a wholesaling context”, *International Journal of Production Economics*, vol. 127, no. 1, pp. 103–111, 2010.

A. A. Syntetos et J. E. Boylan, “On the bias of intermittent demand estimates”, *International Journal of Production Economics*, vol. 71, no. 1-3, pp. 457–466, 2001.

——, “The accuracy of intermittent demand estimates”, *International Journal of Forecasting*, vol. 21, no. 2, pp. 303–314, 2005.

A. A. Syntetos, J. E. Boylan, et J. Croston, “On the categorization of demand patterns”, *Journal of the Operational Research Society*, vol. 56, no. 5, pp. 495–503, 2005.

A. A. Syntetos, K. Nikolopoulos, et J. E. Boylan, “Judging the judges through accuracy-implication metrics : The case of inventory forecasting”, *International Journal of Forecasting*, vol. 26, no. 1, pp. 134–143, 2010.

A. A. Syntetos, R. H. Teunter *et al.*, *On the calculation of safety stocks*. University of Groningen, Faculty of Economics and Busisness, 2014.

A. A. Syntetos, Z. Babai, J. E. Boylan, S. Kolassa, et K. Nikolopoulos, “Supply chain forecasting : Theory, practice, their gap and the future”, *European Journal of Operational Research*, vol. 252, no. 1, pp. 1–26, 2016.

L. J. Tashman et J. M. Kruk, “The use of protocols to select exponential smoothing procedures : A reconsideration of forecasting competitions”, *International Journal of Forecasting*, vol. 12, no. 2, pp. 235–253, 1996.

J. W. Taylor, “Short-term electricity demand forecasting using double seasonal exponential smoothing”, *Journal of the Operational Research Society*, vol. 54, no. 8, pp. 799–805, 2003.

R. H. Teunter et L. Duncan, “Forecasting intermittent demand : a comparative study”, *Journal of the Operational Research Society*, vol. 60, no. 3, pp. 321–329, 2009.

R. H. Teunter, A. A. Syntetos, et M. Z. Babai, “Intermittent demand : Linking forecasting to inventory obsolescence”, *European Journal of Operational Research*, vol. 214, no. 3, pp. 606–615, 2011.

L. Tiacci et S. Saetta, “An approach to evaluate the impact of interaction between demand forecasting method and stock control policy on the inventory system performances”, *International Journal of Production Economics*, vol. 118, no. 1, pp. 63–71, 2009.

R. Tibshirani, “Regression shrinkage and selection via the lasso”, *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

K. H. Van Donselaar, V. Gaur, T. Van Woensel, R. A. Broekmeulen, et J. C. Fransoo, “Ordering behavior in retail stores and implications for automated replenishment”, *Management Science*, vol. 56, no. 5, pp. 766–784, 2010.

B. Van Roy, D. P. Bertsekas, Y. Lee, et J. N. Tsitsiklis, “A neuro-dynamic programming approach to retailer inventory management”, dans *Proceedings of the 36th IEEE Conference on Decision and Control*, vol. 4. IEEE, 1997, pp. 4052–4057.

E. Van Wingerden, R. J. I. Basten, R. Dekker, et W. Rustenburg, “More grip on inventory control through improved forecasting : A comparative study at three companies”, *International Journal of Production Economics*, vol. 157, pp. 220–237, 2014.

M. Wagner, Harvey, *Principles of management science*. Prentice-Hall, 1970.

P. Wallström et A. Segerstedt, “Evaluation of forecasting error measurements and techniques for intermittent demand”, *International Journal of Production Economics*, vol. 128, no. 2,

pp. 625–636, 2010.

D. Waters, *Inventory control and management*. John Wiley & Sons, 2008.

R. Wilson, *A scientific routine for stock control*. Harvard Univ., 1934.

P. R. Winters, “Forecasting sales by exponentially weighted moving averages”, *Management Science*, vol. 6, no. 3, pp. 324–342, 1960.

Y. Zhang et Z. Wang, “A markov decision process model for inventory control under invisible stock loss and inaccurate record”, dans *2017 11th Asian Control Conference (ASCC)*. IEEE, 2017, pp. 2478–2483.