Université de Montréal

Prédiction du délai d'attente en temps réel et modélisation des durées de service dans les centres d'appels multi-compétences

par
Mamadou Thiongane

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Août, 2016

RÉSUMÉ

Dans cette thèse, nous commençons par l'étude de la prédiction de délai d'attente des clients dans les centres d'appels multi-compétences. Le temps d'attente a un impact important sur la qualité du service perçue par les clients. L'annonce du délai d'attente permet de réduire l'incertitude du client à propos de son délai d'attente. Elle peut également augmenter la satisfaction du client et réduire le nombre d'abandons. Ceci nécessite d'avoir un bon prédicteur de délai. Malheureusement les prédicteurs existants ne sont pas adaptés pour les centres d'appels multi-compétences.

Nous proposons trois types de prédicteurs qui utilisent l'apprentissage machine : le premier utilise la régression par les splines cubiques, le second emploie les réseaux de neurones artificiels, et le dernier utilise le krigeage stochastique. Les prédicteurs prennent en entrée le temps d'attente du dernier client de même type à entrer en service, la période d'arrivée du nouveau client, le nombre d'agents des groupes, la longueur de la file des clients de même type, et les longueurs des files d'attente des types servis par les mêmes agents. Ces prédicteurs donnent de bons résultats pour les systèmes multi-compétences, mais un inconvénient est qu'ils ont un grand nombre de paramètres qui doivent être appris à l'avance durant une phase d'entraînement du modèle qui nécessite une grande quantité de données et temps de calcul.

Nous proposons ensuite deux nouveaux prédicteurs de délai qui sont très simples à mettre en œuvre, requièrent peu d'effort d'optimisation, ne nécessitent pas de données, et qui sont applicables dans les centres d'appels multi-compétences. Ils sont basés sur l'historique des temps d'attente des clients. Le premier estime le délai d'un nouveau client en extrapolant l'historique des attentes des clients actuellement dans la file d'attente, en plus du délai du dernier qui a commencé le service, et en prenant une moyenne pondérée. Le second retourne une moyenne pondérée des délais des anciens clients de la même classe qui ont trouvé la même longueur de file d'attente quand ils sont arrivés.

Ensuite, nous nous intéressons à la modélisation des durées de service dans les centres d'appels. En général, les modèles de file d'attente d'Erlang standard sont utilisés pour analyser les opérations dans les centres d'appels. Dans ces modèles, les temps de service des agents sont modélisés comme des variables aléatoires exponentielles indépendantes, identiquement distribuées et de moyenne constante. Plusieurs travaux récents ont montré que la distribution des temps de service est : dépendante du temps, log-normale plutôt qu'exponentielle, et dépend aussi de l'agent.

Nous proposons une modélisation plus réaliste des temps de service dans les centres d'appels qui prennent en compte plusieurs propriétés observées dans les données réelles. Nos modèles prennent en compte : l'hétérogénéité des agents, la dépendance du temps, les corrélations sérielles entre les temps de service d'un agent pour le même type d'appel, et les corrélations croisées entre plusieurs types d'appels servis par le même agent. Nous avons montré que ces modèles prédisent les moyennes des temps de service des agents mieux que les modèles de références considérés. Par la suite, nous montrons par la simulation que ces modèles plus réalistes conduisent à des prédictions des performances du système significativement différentes de celles des modèles de références, et les décisions que pourraient prendre le gestionnaire en observant ces données peuvent mener à des économies de coûts importants dans la pratique.

Mots clés: temps d'attente, apprentissage machine, historique de délai, modélisation, temps de service, simulation

ABSTRACT

In this thesis, we begin with the study of delay prediction of customers in multiskill call centers. Waiting time has an important impact on the quality of service experienced by customers. Delay announcement can reduce customer uncertainty about its delay time. It also can increase customer satisfaction and reduce the number of abandonments. This requires having a good delay predictor. Unfortunately existing predictors are not adapted for multiskill call centers.

We propose three types of predictors that use machine learning: the first uses regression cubic splines, the second employs artificial neural networks, and the latter uses the stochastic kriging. The predictors take as inputs the delay of the last customer of the same type to enter service, the arrival period of the new customer, the staffing of agents groups, the queue length of the same type, and the queue lengths of types served by the same agents. These predictors work well for multiskill call centers, but one drawback is that they have a large number of parameters that must be learned in advance during the training phase that requires a large amount of data and computional time.

We also propose two new delay predictors that are very simple to implement, require little optimization effort, do not need any data, and are applicable in multiskill call centers. They are based on the wait times of previous customers of the same class. The first one estimates the delay of a new customer by extrapolating the wait history of customers currently in queue, plus the delay of last one that started service, and taking a weighted average. The second one takes a weighted average of the delays of the past customers of the same class that have found the same queue length when they arrived.

Next in this thesis, we are also interested in modelling service time in call centers. In general, the standard Erlang queueing models are used to analyze call centers operations. In these models, agent service times are modelled as independent and identically distributed exponential random variables with a constant mean. Several recent studies have shown that the distribution of service time is:

time-dependent, lognormal rather than exponential, and distinct by agent.

We propose a more realistic modelling of service times in call centers that takes into account multiple properties observed in real life data. Our models take into account: the heterogeneity of agents, the time dependence, serial correlation between service time of an agent for the same call type, and the cross-correlations between several call types served by the same agent. We show that these models predict agent average service time better than the considered benchmark models. Thereafter, we show by simulation that these more realistic models lead to system performance predictions significantly different from those of the benchmark models, and decisions that manager could take by observing this data can lead to important cost savings in practice.

Keywords : wait time, machine learning, delay history, modelling, service time, simulation

TABLE DES MATIÈRES

RÉSUMÉ . . . . . **iii**

ABSTRACT . . . . . **v**

TABLE DES MATIÈRES . . . . . **vii**

LISTE DES TABLEAUX . . . . . **xii**

LISTE DES FIGURES . . . . . **xvi**

LISTE DES SIGLES . . . . . **xix**

REMERCIEMENTS . . . . . **xxi**

CHAPITRE 1 : INTRODUCTION **1**

1.1 Les centres d'appels 1

1.2 La gestion des centres d'appels 3

1.3 La prédiction des délais d'attente 4

1.4 La modélisation et la simulation des durées de service 6

1.5 Contributions principales de la thèse 8

1.6 Le plan de la thèse 11

CHAPITRE 2 : DESCRIPTION D'UN CENTRE D'APPELS ET MESURES DE PERFORMANCES **14**

2.1 Description du modèle de centres d'appels 14

2.2 Les mesures de performances 17

CHAPITRE 3 : REVUE DE LA LITTÉRATURE SUR LA PRÉVISION DES DÉLAIS **21**

3.1 L'effet de l'annonce du délai d'attente 21

3.2 La prédiction du délai d'attente des clients 33

3.2.1 Les prédicteurs pour les systèmes avec un seul type de clients et une seule file d'attente 34
3.2.2 Les prédicteurs pour les systèmes multi-compétences 42

CHAPITRE 4 : PRÉDICTEURS DE DÉLAIS POUR LES CENTRES D'APPELS MULTI-COMPÉTENCES BASÉS SUR L'APPRENTISSAGE MACHINE 46

4.1 Introduction 46
4.1.1 Objectifs 46
4.1.2 Le plan du chapitre 48
4.2 Les prédicteurs de délai 48
4.2.1 Approximation de l'espérance conditionnelle du délai 48
4.2.2 Régression par des Splines de lissage (RS) 50
4.2.3 Les réseaux de neurones artificiels (ANN) 50
4.2.4 Le krigeage stochastique (SK) 52
4.3 Expériences numériques 53
4.3.1 Modèles à file unique avec des agents homogènes et durées de service exponentielles 55
4.3.2 Modèles à file unique avec des agents hétérogènes et des durées de service exponentielles 58
4.3.3 Modèles à file unique avec des agents hétérogènes et des durées de service de loi log-normale 62
4.3.4 Modèles N 64
4.3.5 Expériences avec un grand centre d'appels basé sur des données réelles (HQ) 69
4.4 Impact de l'ajout de la période et du nombre d'agents des groupes dans la définition de l'état du système 72
4.4.1 Modèle N avec longues files 73
4.4.2 Le modèle N avec courtes files 73

4.4.3 Exemple avec un grand centre d'appels basé sur des données réelles . . . 74
4.5 La Robustesse des prédicteurs . . . 75
4.5.1 Variation du taux d'arrivée avec le modèle M/M/s+M . . . 77
4.5.2 Variation du taux d'arrivée avec le modèle N . . . 80
4.5.3 La précision des prédicteurs RS, ANN et SK pour les types appels rares . . . 83
4.6 Comparaison des nouveaux prédicteurs avec Q-Lasso . . . 85

CHAPITRE 5 : NOUVEAUX PRÉDICTEURS DE DÉLAI BASÉS SUR L'HISTORIQUE POUR LES SYSTÈMES DE SERVICE . . . 89

5.1 Introduction . . . 89
5.1.1 Contexte et problème . . . 89
5.1.2 Objectifs . . . 90
5.1.3 Organisation du chapitre . . . 91
5.2 Les prédicteurs de délais . . . 92
5.2.1 LES extrapolé (E-LES) . . . 92
5.2.2 Moyenne des LES conditionnelles à la longueur de la file d'attente (AvgC-LES) . . . 94
5.3 Les résultats des simulations . . . 96
5.3.1 Une file d'attente avec un unique type M/M/s+M . . . 97
5.3.2 Une file d'attente unique de type M/LN/s+M . . . 99
5.3.3 Modèle N de centre d'appels . . . 101
5.3.4 Un grand centre d'appel basé sur des données réelles . . . 107
5.4 Conclusion . . . 110

CHAPITRE 6 : PRÉDICTEURS QL POUR LES CENTRES D'APPELS MULTI-COMPÉTENCES ET PRÉDICTION DE LA DISTRIBUTION CONDITIONNELLE DU

DÉLAI D'ATTENTE 112
6.1 Introduction 112
6.2 Nouveaux prédicteurs QL pour les centres d'appels multi compétences 112
6.2.1 Méthodes d'estimation du nombre d'agents des groupes . . . 115
6.2.2 Exemple numérique du modèle N sans abandon 118
6.2.3 Variation de $\tilde{s}$ en fonction du routage 123
6.3 Prédiction de la distribution conditionnelle des temps d'attente . . . 124

CHAPITRE 7 : MODÉLISATION DES DURÉES DE SERVICE DANS LES CENTRES D'APPELS 126
7.1 Introduction 126
7.2 Revue de littérature 127
7.2.1 Hétérogénéité des agents 128
7.2.2 Dépendances entre les temps de service 129
7.2.3 Dépendances avec le temps 130
7.2.4 Distribution Log-normale 130
7.3 Analyse préliminaire des données 131
7.3.1 Vue d'ensemble 132
7.3.2 Statistiques sur les temps de service 133
7.3.3 La cohorte C de 200 agents 142
7.4 Modèles de Temps de Service 145
7.4.1 Les Modèles de benchmark B1 et B2 145
7.4.2 Modèle A1 147
7.4.3 Modèle A2 : Corrélations sérielles 148
7.4.4 Modèle A3 : Corrélations sérielle et croisée 152
7.5 Qualité de l'ajustement des modèles 154
7.5.1 Modèle des résidus 154
7.6 Prédictions de la moyenne des temps de service 156
7.6.1 Prédictions de deux semaines en avant 159
7.6.2 Prédictions d'une journée en avant 159

7.6.3 Proportion de victoires pour chaque modèle 162
7.7 Simulation 163
7.7.1 Estimations des paramètres 164
7.7.2 Impact des différents modèles de temps de service sur les performances 168
7.7.3 Impact de la sélection d'agents 172
7.8 Conclusion et remarques 177

CHAPITRE 8 : CONCLUSION 185

BIBLIOGRAPHIE 189

LISTE DES TABLEAUX

4.1 Les performances des différentes modèles M/M/s utilisés. . . . 56
4.2 Les RRASEs pour le modèle M/M/s. 56
4.3 Les performances des différents modèles M/M/s+M utilisés. . 57
4.4 Les RRASEs pour le modèle M/M/s+M. 58
4.5 Le taux de service et moyenne du temps de service des 12 agents pour le type d'appel A. 59
4.6 Les mesures de performances pour les modèles M/M/s+M . . 59
4.7 Les RRASEs des centres d'appels C1, C2 et C3 avec les prédicteurs QL(Méthode 1) et QL(Méthode 2). 61
4.8 Les RRASEs pour les modèles M/M/s+M. 62
4.9 La moyenne m et la variance v des temps de service des agents pour le type d'appel B. 63
4.10 RRASE des prédicteurs pour le modèle M/LN/10+M avec des taux d'arrivée variable dans le temps. 64
4.11 Le RRASE pour le N. modèle avec de courtes files. 66
4.12 Les RRASEs pour le modèle N avec longues files. 68
4.13 Mesures de performances moyennes pour le grand exemple. . . 70
4.14 RRASEs pour les 6 types d'appels du grand exemple. 72
4.15 Les RRASEs pour le modèle N avec longues files. 73
4.16 Le RRASE pour le N. modèle avec de courtes files. 74
4.17 RRASEs pour les 6 types d'appel de l'exemple basé sur des données réelles. 74
4.18 Performances du modèle M/M/s+M avec les variations du taux d'arrivée 78
4.19 RRASE M/M/s+M avec variation du taux d'arrivée 79
4.20 RRASE des prédicteurs pour un centre d'appels M/M/s+M avec "busyness factor". 80
4.21 RRASE du modèle N avec variation du taux d'arrivée 81

4.22 Les RRASEs pour le modèle N avec données d'entraînement et de test collectées sur plusieurs journées. 83
4.23 Mesures de performances moyennes pour le modèle W.. 84
4.24 Les RRASEs pour le modèle W. 85
4.25 Le RRASE pour le N. modèle avec de courtes files. 86
4.26 Les RRASEs pour le modèle N avec longues files. 86
4.27 Le RRASE des prédicteurs pour les 6 types d'appels du centre d'appels HQ. 87

5.1 RRASEs for the M/M/20+M example. 98
5.2 RRASE des prédicteurs pour le modèle M/M/10+M avec des taux d'arrivée variable dans le temps. 100
5.3 RRASE pour chaque type d'appel, pour le modèle N avec courtes files. 102
5.4 Mesures de performance moyenne du modèle N courtes files avec agents hétérogènes et temps de service de loi log-normale. 104
5.5 RRASE pour chaque type d'appel, pour le modèle N avec courtes files et des agents hétérogènes. 105
5.6 RRASE pour chaque type d'appel, pour l'exemple du modèle N.105
5.7 Mesures de performance moyennes pour l'exemple du modèle N.107
5.8 RRASE pour chaque type d'appel, pour l'exemple du modèle N.107
5.9 RRASEs for the 6 call types of the larger example. 108

6.1 Les valeurs de $\tilde{s}$ pour $k = 0$ à $k = 20$ pour QL1 120
6.2 RRASE×100 pour le modèle N sans abandon avec une seule période. 122
6.3 RRASE modèle N avec une augmentation de 1% 122
6.4 RRASE modèle N avec une augmentation de 2% 123

7.1 Résultats pour le Modele $A2$ pour 3 différentes combinaisons agent/type d'appel. Les estimations ponctuelles des coefficients du modèle sont montrées avec les erreurs standard et p-valeurs pour des significations statistiques des t-tests. 153
7.2 Résultats pour le Modele $A3$ pour l'agent i_0, présenté dans le tableau 7.1, répondant à 3 different types d'appels, numérotés de 1 à 3. Les estimations ponctuelles des coefficients du modèle sont montrés avec les erreurs standard et p-valeurs indiquant ce qui est statistiquement significatif. 155
7.3 Résumé statistique du carré des résidus avec chaque modèle, à travers la cohorte de C agents. 159
7.4 Précision des prédictions pour les Modèles $A1$, $A2$, et $A3$, moyenne à travers la cohorte de C agents. 162
7.5 Les proportions où un modèle donné est gagnant, c.-à-d., donne les plus petites mesures de performances, à travers la cohorte de C agents. 163
7.6 Les valeurs de v_k estimé avec la méthode 1, e_k, σ_ν^2 , et E[1/N] pour certains agents au vendredi de la semaine 45 pour le modèle A3. 168
7.7 Les valeurs de v_k estimé avec la méthode 2, e_k, σ_ν^2 , et E[1/N] pour certains agents au vendredi de la semaine 45 pour le modèle A3 169
7.8 Moyennes des temps de service observées M et prédites $\hat{M}$, σ^2 et σ_γ^2 pour certains agents au vendredi de la semaine 45 pour chaque modèle. 171
7.9 RMSE et MAPE des erreurs de prédictions pour le type d'appel F au vendredi de la semaine 45. 171
7.10 RMSE et MAPE des erreurs de prédictions pour le type d'appel E au vendredi de la semaine 45. 171

7.11 Performances estimées et intervalles de confiance pour notre modèle N. 173
7.12 Performance estimée et intervalles de confiance pour notre modèle N, avec des agents rapides. 173
7.13 Performance estimée et intervalles de confiance pour notre modèle N avec des agents EF lents. 176
7.14 Staffing des périodes les agents F 176
7.15 Paramètres de forme et d'échelle des périodes pour la distribution Gamma dans le processus d'arrivée pour type d'appel F. 178
7.16 Paramètres de forme et d'échelle des périodes pour la distribution Gamma dans le processus d'arrivée pour type d'appel E. 179
7.17 Matrice de corrélation entre les taux d'arrivée pour le type d'appel F partie 1. 181
7.18 Matrice de corrélation entre les taux d'arrivée pour le type d'appel F partie 2. 182
7.19 Matrice de corrélation entre les taux d'arrivée pour le type d'appel E partie 1. 183
7.20 Matrice de corrélation entre les taux d'arrivée pour le type d'appel E partie 2. 184

LISTE DES FIGURES

2.1 Les modèles multi-compétences V, N et W 17

4.1 Modèle N avec courtes files : Distribution du temps d'attente des clients qui ont attendu et reçu le service, pour chaque type. 66

4.2 Modèle N avec courtes files : Distribution des erreurs de prédiction (délai estimé moins délai réel) pour le type 1 et type 2. 67

4.3 Modèle N avec longues files : Distribution du temps d'attente pour les clients qui ont attendu et servi. 68

4.4 Modèle N avec longues files : Distribution des erreurs de prédiction (délai estimé moins délai réel) pour les types d'appels 1 et 2. 69

4.5 RRASE des prédicteurs pour modèle M/M/s+M en fonction du taux d'arrivée des données de test. 78

4.6 RRASE des prédicteurs pour le modèle N en fonction du taux d'arrivée $\tilde{\lambda}_1$, type 1. 81

4.7 RRASE des prédicteurs pour le modèle N en fonction du taux d'arrivée $\tilde{\lambda}_2$, type 2. 82

5.1 Modèle M/M/s+M : distribution de l'erreur de prédiction. . . 98

5.2 Jour 1 99

5.3 Jour 2 99

5.4 Jour 3 100

5.5 Jour 4 100

5.6 Modèle N avec courtes files : Distribution des erreurs de prédictions (délai estimé moins délai réel) pour le type 1 et le type 2. 103

5.7 Modèle N avec longues files : Distribution des erreurs de prédiction (délai estimé moins réel) pour les types 1 et 2. 106

5.8 Modèle HQ : Distribution des erreurs de prédiction pour les types 1 et 2. 108
5.9 Modèle HQ : Distribution des erreurs de prédiction pour les types 3 et 4. 109
5.10 Modèle HQ : Distribution des erreurs de prédiction pour les types 5 et 6. 109

6.1 Système alternatif du modèle N. 113
6.2 Variation de $\tilde{s}$ en fonction k pour le type d'appel 1. 121
6.3 Variation de $\tilde{s}$ en fonction k pour le type d'appel 2. 121
6.4 QL1, Variation de $\tilde{s}$ en fonction du routage pour $k = 0$ à $k = 20$.123

7.1 Nombre moyen d'agents par semaine et les bandes de confiance à 95% correspondant. 133
7.2 Nombre moyen d'appels répondus et les bandes de confiance à 95% correspondant. 134
7.3 Chaque point correspond à une paire (moyenne, variance) pour type donné. 135
7.4 Moyenne du temps de service pour différents agents traitant le type d'appel A en fonction du nombre total d'appels répondus par année. La ligne horizontale est la moyenne globale pour tous les agents. 137
7.5 Moyenne du temps de service pour différents agents traitant le type d'appel B en fonction du nombre total d'appels répondus par année. La ligne horizontale est la moyenne à travers tous les agents. 138
7.6 Les variances du temps de service estimées pour les agents traitant le type d'appel A en fonction du nombre total d'appels répondus par année. 139

7.7 Les variances du temps de service estimées pour les agents traitant le type d'appel B en fonction du nombre total d'appels répondus par année. 140
7.8 La moyenne des temps de service pour 4 agents traitant le type d'appels B versus indice de la journée. 142
7.9 La moyenne journalière des temps de service pour un agent traitant de multiples types d'appels et dont la liste des compétences augmente au jour 208. 143
7.10 L'évolution de la moyenne des temps de service de l'agent a_1 pour le type d'appels A, et le meilleur ajustement linéaire. . . 144
7.11 Diagramme Q-Q des résidus du Modèle A1 pour l'agent a_1 et les bandes de confiance à 95%. 149
7.12 Diagramme Q-Q des résidus du Modèle A1 pour l'agent a_2 et les bandes de confiance à 95%. 150
7.13 Boîte de moustaches du RMSE du modèle des résidus lors de l'ajustement de tous les modèles aux données de la cohorte de C agents. 157
7.14 ECDF pour le RMSE des modèles de résidus lors de l'ajustement aux données à la cohorte de C agents. 158
7.15 ECDF pour le RMSE pour les prévisions d'une journée à l'avance, à travers tous les agents de la cohorte C. 160
7.16 ECDF pour le MAPE pour les prévisions d'une journée à l'avance, à travers tous les agents de la cohorte C. 161
7.17 Histogramme du temps service moyen et niveau de service pour les types d'appels 1 et 2 avec tous les modèles. 174
7.18 Histogrammes du temps de service moyen et du niveau de service pour les types d'appels 1 et 2 avec tous les modèles. . . . 175
7.19 Histogrammes du temps de service moyen et du niveau de service pour les types d'appels 1 et 2 avec tous les modèles. . . . 177

LISTE DES SIGLES

ANN	Prédicteur qui utilise les réseaux de neurones artificiels.
AQS	La longueur moyenne de la file d'attente.
ASE	La valeur empirique du MSE.
AvgC-LES	Prédicteur qui utilise la moyenne les temps d'attente des clients passés de même type qui ont trouvé la même longueur de file d'attente quand ils sont arrivés.
AWT	"Acceptable Waiting Time" ou temps d'attente acceptable
DH	"Delay History predictor" ou Prédicteur de délai qui utilise l'historique du système.
E-LES	Prédicteur qui utilise la moyenne pondérée des temps d'attente extrapolés des clients déjà dans la file, plus le dernier qui est entré en service.
FCFS	"First-Come First-Served" ou premier arrivé premier servi.
HOL	L'estimateur qui estime le délai d'attente du nouveau client par le délai d'attente enregistré jusqu'ici par le client à la tête de la file.
i.i.d	Indépendant et identiquement distribué.
LES	Le prédicteur qui estime le délai d'attente du nouveau client par le délai d'attente du dernier client à entrer en service.

MAPE "Mean absolute percentage error" ou la moyenne absolue du pourcentage d'erreur.

MSE Erreur quadratique moyenne.

NI Le prédicteur qui estime le délai d'attente du nouveau client par le temps d'attente moyen du système.

PA La proportion des clients qui ont abandonné.

PD La proportion des clients qui ont attendu.

QL "Queue Length Predictor" ou ou Prédicteur de délai qui utilise la longueur de la file d'attente et les paramètres du système.

RCS Le prédicteur qui estime le délai d'attente du nouveau client par le délai d'attente enregistré par le dernier client parmi ceux qui ont récemment terminé leur service

RASE La racine carrée du ASE

RRASE Le ASE normalisé par la moyenne des temps d'attente des clients qui ont eu un temps d'attente strictement positif.

RS Prédicteur qui utilise la régression par les splines de lissage.

SK Prédicteur qui utilise le krigeage stochastique.

SL Niveau de service.

REMERCIEMENTS

Je remercie infiniment mon directeur de recherche, Pierre L'Ecuyer. Je le remercie de m'avoir proposé ce sujet de recherche et de m'avoir soutenu financièrement. Ce sont ces conseils, critiques, suggestions, et idées qui m'ont permis de réaliser ce travail.

Je remercie également le programme de bourse de la francophonie de m'avoir financé pendant 4 ans. Un grand merci à toute l'équipe de gestion. Je remercie particulièrement Wyean Chan avec qui j'ai beaucoup collaboré durant cette thèse. Je remercie l'ensemble de mes collègues du Département, Shohre Zehtabian, Nazim Régnard, David Munger, Richard Simard, Anh Ta et Amal Boukhdhir pour leur soutien. Je remercie les co-auteurs d'un des trois articles Rouba Ibrahim et Haipeng Shen.

Je remercie mes parents et ma femme Khady Camara pour leur soutien durant toutes ces longues années. Je remercie tous mes amis, frères et soeurs plus particulièrement Barham Thiam, Ousseynou Diop, Aliou Ndao, Pathé Ndome, Mouhamed Ndiaye, Mamadou Wade, Mame Astou Biteye, Ndeye Marem Fall, Ousmane Diagne, Abdalla Ndiagne, Ousmane Sow, Mouhamed Ndong, Mbaye Ndoye, Alioune Fall, Rafik Gouiaa pour leur soutien et leur encouragement.

CHAPITRE 1

INTRODUCTION

Un centre d'appels se définit comme étant un ensemble de ressources, généralement du personnel, des ordinateurs et des équipements de télécommunication, qui permettent d'offrir des services par téléphone (Gans et al., 2003). Les centres d'appels sont aujourd'hui les éléments clés de presque toutes les grandes organisations. Ils sont utilisés, par exemple, pour fournir de l'information ou du support aux clients. Ils permettent aux compagnies de faire des ventes et aux clients de payer leurs factures par téléphone, etc. Ils sont aussi utilisés par les services d'information gouvernementaux et les services d'urgence (police, ambulance), etc. Il y a des milliers de centres d'appels dans le monde, avec des tailles en termes d'employés allant d'un à plusieurs milliers. Actuellement, le nombre d'employés dans les centres d'appels dépasse les 4 millions de personnes pour seulement les États-Unis et le Canada.

1.1 Les centres d'appels

Les appels traités dans un centre d'appels sont en général classifiés en deux catégories selon l'origine de l'appel. Les appels émis par les clients et reçus par les fournisseurs de service sont appelés *appels entrants* ou "*inbounds calls*", et les appels du fournisseur de service vers les clients sont appelés *appels sortants* ou "*outbound calls*". Les centres d'appels qui traitent à la fois les deux catégories d'appels sont appelés les centres d'appels mixtes ou "*blended call centers*". Les centres d'appels ont aujourd'hui évolué en centres de contacts en offrant des services par courriels, chat, fax, etc. Dans cette thèse, nous nous concentrons sur les centres d'appels avec des appels entrants. Dans ces centres, les agents qui traitent les appels sont souvent appelés *les représentants du service à la clientèle* ou "*customer service representatives*" (CRS) ou encore plus souvent "agents".

Nous allons maintenant décrire le principe de fonctionnement des centres d'appels. Les appels émis par les clients en général aboutissent à un système automatique qui est appelé "Interactive Voice Response" (IVR) à travers un commutateur appelé PABX ("*Private Automatic Branch eX-change*"). L'IVR permet de recueillir les informations sur le client et de déterminer le type de service désiré par ce dernier. En d'autres termes, l'IVR permet d'identifier le client et de déterminer le type d'appel. Le client interagit avec l'IVR en utilisant les touches du clavier de son téléphone ou la voix. Pour certains types de service, l'IVR peut offrir aux clients la possibilité de s'auto servir. Gans et al. (2003) indiquent que dans les banques 80% des clients utilisent l'IVR pour s'auto servir.

Dans le cas où le client communique son désir de parler à un agent, l'IVR transmet l'appel à un *distributeur automatique d'appel* ou "*automatic call distributor*" (ACD). L'ACD est un routeur spécialisé sur lequel la politique de routage des appels est implémentée. Les ACDs modernes sont très sophistiqués et permettent la programmation de politiques de routage souvent très complexes. L'ACD affecte l'appel à un agent disponible ayant la compétence pour le traiter. S'il n'y a aucun agent disponible pour le servir, l'appel est mis attente dans une file d'attente. Les centres d'appels disposent d'un système de files d'attente qui regroupe les appels en attente de service. Durant cette période d'attente, les clients impatients peuvent raccrocher leur téléphone. On parle dans ce cas d'abandon. Pour occuper les clients pendant cette période d'attente, certains fournisseurs de service mettent souvent de la musique qui est accompagnée d'un message d'excuse pour l'attente qu'ils auront à faire avant la disponibilité d'un agent, ou bien fournissent des informations sur les nouveaux produits et services. Actuellement pour diminuer l'incertitude à propos du délai d'attente et augmenter la satisfaction des clients, certains centres d'appels informent aussi les clients sur leurs délais d'attente prédits. D'autres vont plus loin encore, en proposant une option de rappel si le temps d'attente prédit est jugé trop élevé (par exemple un temps d'attente estimé à 45 minutes). D'ailleurs le développement de méthodes de prédiction de délais adaptées aux centres d'appels multi-compétences est l'un des sujets de cette thèse.

Une fois connecté à un client par l'ACD, l'agent peut parler au client par téléphone et en même temps travailler sur un terminal. Le terminal permet à l'agent d'accéder à un serveur de données des clients ou "*customer data server*" qui contient les informations sur les clients. Le lien entre le "*customer data server*" et l'ACD est géré par un "*middleware*" appelé "*customer-telephone integration*" (CTI), en montrant par exemple le dossier du client par son numéro de téléphone.

Notons aussi que c'est au niveau de l'ACD que les données statistiques du centre d'appels (tels que les temps d'arrivées, les abandons, les durées de service, la longueur de la file d'attente, les délais d'attente, etc.) sont collectées. Pour plus de détails sur le fonctionnement des centres d'appels, voir Avramidis et L'Ecuyer (2005), Chan (2013), Gans et al. (2003), Pichitlamken et al. (2003) et Koole (2013).

1.2 La gestion des centres d'appels

Les centres d'appels sont devenus aujourd'hui très complexes et les modèles mathématiques traditionnels d'Erlang ne sont plus adaptés pour leur modélisation. Les centres d'appels simples ont évolué maintenant en centres d'appels plus complexes appelés centres d'appels multi-compétences. Un centre d'appels téléphoniques multi-compétences est un centre d'appels téléphoniques qui reçoit plusieurs types d'appels, où différents types d'agents avec diverses combinaisons de compétences doivent gérer différents types d'appels qui arrivent au hasard au centre d'appels. Chaque agent peut généralement gérer uniquement un sous-ensemble de types d'appels. Chaque type d'appel exige une compétence particulière et chaque groupe d'agents a un sous-ensemble donné de ces compétences, de sorte que les agents de ce groupe peuvent gérer les appels qui ont besoin de ces compétences et seulement ces compétences.

La gestion efficace ces centres d'appels est devenue une tâche très difficile pour les gestionnaires. Il y a beaucoup de sources d'incertitude à gérer. Parmi les sources d'incertitude, nous pouvons citer les taux d'arrivées des appels qui sont généralement des processus stochastiques ou doublement stochastiques. Nous pouvons aussi

citer, les temps de service des appels qui sont aléatoires et dont la distribution peut dépendre du type d'appel et de l'agent qui traite l'appel.

Le temps d'attente d'un client dans ces systèmes est très difficile à déterminer. En effet, pour certains centres d'appels, la plupart des appels à leur arrivée trouvent que tous les agents ayant les compétences pour les traiter sont occupés. Ces appels sont stockés dans des files d'attente (invisibles aux clients) jusqu'à ce qu'un agent ayant les compétences pour les traiter soit disponible pour les servir. Du fait de la complexité des modèles actuels et du routage, le temps qu'un client doit attendre peut dépendre de la longueur de sa file d'attente, des longueurs actuelles et futures des autres files d'attente, du staffing des groupes d'agents, de la période de la journée, etc. Pour plus détails sur les problèmes des centres d'appels, voir Akşin et al. (2007a, b), Gans et al. (2003), Koole et Mandelbaum (2002) et Koole (2013).

Dans cette thèse, nous avons travaillé sur deux problèmes des centres d'appels multi-compétences. Le premier est la prédiction de délai d'attente des clients en temps réel. Le second est la modélisation et simulation des durées de service des agents.

1.3 La prédiction des délais d'attente

Nous nous intéressons dans ce travail à développer et étudier des méthodes pour estimer le temps d'attente d'un client lors de son arrivée au centre d'appels dans le but éventuel d'annoncer cette information (ou une partie) au client. On peut aussi utiliser cette information pour faire une éventuelle proposition de rappel au client au cas où le temps d'attente estimé est supérieur à un certain seuil (par exemple 30 minutes). Ce seuil peut être interprété (par le gestionnaire du centre d'appels) comme étant le délai d'attente au-delà duquel la probabilité que le client quitte la file avant d'être servi élevée, ou bien être interprété comme une congestion du centre d'appels. Le client pourra ainsi prendre une décision plus éclairée pour ou bien attendre, ou bien abandonner, ou bien demander d'être rappelé, etc. Dans le cas où le client choisit d'être rappelé, différents mécanismes de rappel peuvent être

utilisés. Dans certains cas, le rang dans la file d'attente du client qui opte pour le rappel est toujours maintenu. Autrement dit le client quitte réellement la file d'attente, mais son rang dans la file est toujours virtuellement maintenu. Ainsi le rappel du client est effectué quand son tour arrive à la file. Cette stratégie de rappel est la plus utilisée en pratique. Dans d'autres cas, le rappel du client est effectué après une certaine durée. Cette durée peut-être une valeur fixe (elle reste toujours la même pour tous les clients rappelés) ou bien une variable aléatoire dépendant de l'état du système lors de la proposition du rappel.

Plusieurs travaux ont montré que fournir des informations précises sur le délai d'attente aux clients dans les systèmes de file d'attente invisible tels les centres d'appels réduit l'incertitude des clients à propos du temps d'attente et augmentent la satisfaction des clients ; voir par exemple Cleveland et Mayben (1999), Hui et Tse (1996a), Katz et al. (1999), Maister (1984), Munichor et Rafaeli (2007), Taylor (1994a), Whitt (1999a). Cette information fournie peut influencer le comportement des clients et diminuer considérablement le nombre d'abandons et augmenter le taux service, etc. Elle peut aussi aider les gestionnaires à bien gérer leur système. Par exemple si les attentes estimées sont trop longues pour certains types d'appels, le gestionnaire peut réagir par l'augmentation du nombre d'agents des groupes qui traitent ces appels dans un futur proche.

L'estimation du délai d'attente dans les centres d'appels téléphoniques multi-compétences est une question qui intéresse beaucoup les gestionnaires des centres d'appels. L'estimation du temps d'attente peut être sous la forme d'une espérance conditionnelle (moyenne, conditionnelle à l'état actuel du système), ou bien sous la forme d'une densité de probabilité des temps d'attente conditionnelle à l'état courant du système, etc. On pourrait aussi vouloir réestimer le temps d'attente résiduel du client régulièrement pour mettre à jour la prévision. Dans presque tous les cas, les méthodes de prévisions seront des heuristiques simples.

En général, deux familles de prédicteurs de délai d'attente sont considérées. Dans la premiere, nous avons les prédicteurs de délai basés sur l'historique ou "*delay history predictors*" (DH) et dans la seconde famille, nous avons les prédicteurs de

délai basés sur la longueur de la file d'attente ou "*queue length predictors*" (QL). Les prédicteurs DH exploitent l'information sur l'historique récent des délais d'attente des clients déjà servis. Les prédicteurs QL exploitent la connaissance de la longueur de la file (nombre de clients en attente) observée à l'arrivée du client, les paramètres du système comme le taux de service, le taux d'abandon et le nombre de serveurs. Une description plus détaillée de ce qui a déjà été fait sera donnée dans la revue de littérature au chapitre 3.

Nous notons que pratiquement tout ce qui a été fait dans le passé ne s'applique qu'aux centres d'appels avec un seul type d'appel (une seule file) et que cela ne s'applique pas dans le cas des centres d'appels multi-compétences. Les prédicteurs QL ne s'étendent pas naturellement dans le contexte multi-competence. Pour les utiliser dans ce nouveau contexte, il leur faudrait prendre en compte le partage des compétences des agents et la politique de routage, et cela semble compliqué et difficile. Les prédicteurs DH existants peuvent être utilisés dans le cas multi-compétence mais les erreurs de prédiction de ces derniers sont très grandes surtout quand il y a une variation non négligeable dans les processus d'arrivées ou une variation du nombre d'agents dans le temps. Malheureusement, ces variations sont importantes dans les centres d'appels multi-compétences. Il faudra développer des prédicteurs DH qui sont adaptés à ce nouveau contexte.

Dans cette thèse, nous proposons des prédicteurs de délais pour les centres d'appels multi-compétences qui peuvent être classés en deux catégories. La première catégorie de prédicteurs basés sur une approche heuristique, utilise des méthodes de l'apprentissage machine. La seconde catégorie de prédicteurs basés également sur des heuristiques utilise l'historique du système.

1.4 La modélisation et la simulation des durées de service

Dans cette thèse, nous avons travaillé aussi sur la modélisation des durées de service dans les centres d'appels. Habituellement, les modèles de file d'attente d'Erlang standard sont utilisés pour analyser les opérations dans les centres d'appels. Dans

ces modèles, les temps de service des agents sont modélisés comme des variables aléatoires exponentielles indépendantes, identiquement distribuées et de moyenne constante. Beaucoup de travaux récents ont montré qu'au-delà de cette hypothèse standard de modélisation, il y a des conséquences opérationnelles importantes.

Brown et al. (2005), Deslauriers (2003) et Shen et Brown (2006) ont observé que les temps de service ne sont pas exponentiellement distribués, comme on l'a traditionnellement supposé, mais ils sont plutôt distribués suivant une loi log-normale.

L'analyse des données recueillies dans les centres d'appels sur les durées de services des agents a montré que les agents sont hétérogènes. Il y a diverses études théoriques sur les modèles de files d'attente avec des serveurs hétérogènes ; voir par exemple Armony (2005), Armony et Mandelbaum (2011), Armony et Ward (2010), Gurvich et Whitt (2009). Dans ces travaux, les auteurs ont montré que les gestionnaires peuvent prendre en compte l'hétérogénéité des agents lors du routage des appels pour améliorer certaines mesures de performances. Par exemple, router les appels entrants vers les agents libres les plus rapides réduit le temps d'attente des clients.

Aldor-Noiman et al. (2009), Mandelbaum et al. (1999) et Liu et Whitt (2011) ont observé que les temps de service sont dépendants du temps et la présence de taux de service variant dans le temps aura un impact opérationnel non négligeable.

Delasay et al. (2016), Dong et al. (2015) et Feldman et al. (2015) ont observé que les temps de service successifs d'un agent sont souvent dépendants. Il y existe d'autres travaux qui ont étudié l'impact de cette dépendance sur les performances des systèmes. Par exemple Whitt (2002) a étudié cette dépendance dans un système de files d'attente mono serveur et Dong et al. (2012) dans un système de files d'attente multi serveur. La conclusion récurrente de ces études est que la non-considération de cette dépendance a des conséquences opérationnelles importantes.

Dans cette thèse, nous proposons des modèles qui tiennent en compte l'ensemble des propriétés citées ci-dessus. Nos modèles prennent en compte plusieurs propriétés réalistes telles que : l'hétérogénéité de l'agent, la dépendance avec le temps, les corrélations sérielles entre les temps de service d'un agent donné pour un type

d'appel donné, et les corrélations croisées entre plusieurs types d'appels traités par le même agent. Nous comparons nos modèles avec les modèles standards, par exemple, au cas où la moyenne des temps de service ne dépend que du type d'appel. Nous constatons que les modèles qui exploitent les propriétés ci-dessus s'adaptent beaucoup mieux aux données, à la fois dans l'échantillon d'entraînement et en dehors de l'échantillon.

Une investigation empirique des temps de service recueillis au centre d'appels d'Hydro-Québec a été réalisée et des modèles efficaces pour les durées de service ont été proposés. Nous avons comparé à travers plusieurs exemples le pouvoir de prédiction de ces modèles à ceux des modèles de benchmark. Nous avons constaté que les modèles proposés prédisent mieux les moyennes des temps de service que les modèles de benchmark utilisés jusqu'à présent. Par la suite, nous avons montré par la simulation que cette modélisation efficace des temps de service est également importante d'un point de vue opérationnel.

1.5 Contributions principales de la thèse

Dans cette thèse, nous nous sommes intéressés à la prédiction de délais en temps réel des clients dans les centres d'appels en général et dans les centres d'appels multi-compétences en particulier, mais aussi à la modélisation des durées de service dans ces derniers. Les principales contributions de la thèse sont les suivantes. En premier, nous avons développé des prédicteurs de délais qui sont meilleurs que les prédicteurs QL dans les situations réalistes des systèmes à file unique, et qui sont adaptés pour les centres d'appels multi-compétences. Ces prédicteurs utilisent l'apprentissage machine et ils sont publiés en partie dans Thiongane et al. (2015). Plusieurs travaux ont montré que l'information du délai d'attente des clients à leur arrivée au système est utile pour le gestionnaire afin qu'il adopte une stratégie qui peut diminuer considérablement les abandons et augmenter la satisfaction des clients. Sachant que les prédicteurs QL et DH existants ne sont pas adaptés (le premier n'étant pas applicable et le second donne de grandes erreurs de prédiction)

pour prédire le temps d'attente des clients dans les systèmes multi-compétences, nous avons proposé de nouveaux prédicteurs adaptés pour ces systèmes.

Vu la complexité des modèles multi-compétences actuels, nous savons qu'il est difficile de développer des formules mathématiques pour prédire le temps d'attente des clients dans les files d'attente. Une façon d'approcher ses formules c'est d'utiliser les méthodes de l'apprentissage machines. Pour chaque type d'appel k, nous définissons une *fonction de prédiction* $F_{k,\theta}(\boldsymbol{x})$ dépendant de l'état du système $\boldsymbol{x}$ où θ est un vecteur de paramètre à estimer. L'état du système est un vecteur constitué du délai attente du dernier client de type k qui est entré en service, de la longueur de la file d'attente actuelle pour le type d'appel k, des longueurs de file d'attente pour tous les types $i \neq k$ pour lequel il existe un agent qui peut servir les deux types k et i, de la période d'arrivée de l'appel, et du staffing des groupes d'agents. Le vecteur de paramètres qui définit la fonction est "optimisé" (ou *appris*) pour minimiser l'erreur quadratique moyenne de prédiction, en se basant soit sur des données historiques réelles, ou sur des données obtenues à partir d'une simulation du modèle de centre d'appels. Nous avons utilisé trois méthodes. La première méthode utilise la régression par des splines de lissage (de Boor, 1978), la seconde méthode utilise le krigeage stochastique (Ankenman et al., 2010, Staum, 2009), et la troisième méthode utilise les réseaux de neurones artificiels (Bengio et al., 2012, LeCun et al., 2015). Les résultats numériques pour plusieurs exemples montrent que les nouveaux prédicteurs sont meilleurs que les prédicteurs QL dans les centres d'appels réalistes avec une seule file d'attente, un seul type d'appel, des agents hétérogènes, et des durées de service de loi log-normale. Ils sont aussi largement meilleurs que les prédicteurs DH qui utilisent l'historique du système aussi bien dans les systèmes à file unique, que dans les systèmes multi-compétences. Nous avons montré que les prédicteurs sont robustes face à la variation des taux d'arrivée et face à la variation du staffing.

Deuxièmement, nous avons développé des prédicteurs DH qui utilisent l'historique du système pour les systèmes multi-compétences (Thiongane et al., 2016). Ces nouveaux prédicteurs sont attrayants parce qu'ils sont très simples à implémenter

en pratique, possèdent très peu de paramètres, et capturent plus rapidement les changements dans le système que les autres prédicteurs DH existants.

Nous avons proposé deux prédicteurs DH. Le premier prédicteur extrapole le délai d'attente des clients dans la file d'attente. Il estime le délai d'attente du nouveau client par la moyenne des temps d'attente extrapolés et du temps d'attente du dernier client entré en service. Ce prédicteur donne de meilleurs résultats que tous les autres prédicteurs DH aussi bien dans les systèmes avec une seule file que dans les systèmes multi-compétences. Il utilise des informations incomplètes, mais fraîches, donc il devrait capturer plus rapidement les changements du système tel que la variation du taux d'arrivée et du nombre de serveurs. Le second prédicteur, que nous avons proposé, s'inspire du prédicteur QL qui utilise une formule mathématique pour calculer l'espérance du temps d'attente d'un client conditionnelle à la longueur de la file observée à l'arrivée. Notre prédicteur utilise les délais d'attente des clients ayant observé la même longueur de file d'attente pour estimer cette espérance conditionnelle. Il est meilleur que le premier, donne de bons résultats dans les systèmes multi-compétences. Il est aussi très compétitif avec les prédicteurs QL dans les systèmes à file unique avec des temps de service exponentiels et des agents homogènes (où QL est le prédicteur optimal), mais dans les systèmes à file unique avec des agents hétérogènes et des durées de service de loi log-normale (modélisation plus réaliste pour les centres d'appels), ce nouveau prédicteur est largement plus performant que QL.

Ibrahim et al. (2016b) proposent une modélisation des durées de services dans les centres d'appels qui tient compte de l'hétérogénéité des agents, de la dépendance avec le temps des durées de service d'un agent, des corrélations sérielles entre les temps de service d'un agent donné pour un type d'appel donné, et des corrélations croisées entre plusieurs types d'appels traités par le même agent. Notre principale contribution dans ce travail est d'avoir montré par la simulation que les modèles proposés ont un impact opérationnel dans les performances des centres d'appels. Pour y parvenir, nous avons proposé des méthodes pour estimer certains paramètres nécessaires à la simulation des modèles. Nous avons aussi étudié plus en détail le

pouvoir de prédiction des modèles proposés. Une autre contribution importante que nous ne détaillerons pas dans cette thèse est d'avoir développé et intégré un module dans le simulateur des centres d'appels ContactCenters (Buist et L'Ecuyer, 2005) qui a permis de simuler les nouveaux modèles. Le nouveau module ajouté permet de spécifier pour chaque agent i du groupe la liste des distributions de ces durées de service pour les types qu'il peut servir. Nous aurons une distribution pour chaque type d'appel j que cet agent i peut traiter.

1.6 Le plan de la thèse

Le reste de ce document est organisé comme suit. Au chapitre 2, nous faisons une description des modèles de centres multi-compétences sur lesquels nous nous concentrons dans cette thèse. Nous définissons aussi dans ce chapitre l'ensemble des mesures de performances qui sont utilisées tout au long de cette thèse. Au troisième chapitre, nous faisons une revue de la littérature des travaux sur la prédiction de délais dans les systèmes de service. La plupart des travaux présentés dans ce chapitre sont effectués pour les systèmes à file unique. Il existe très peu de travaux effectués pour les systèmes multi-compétences. Au début de ce chapitre, nous avons d'abord visité deux sujets qui motivent la prédiction de délai. Le premier est l'étude des effets de l'annonce du délai d'attente dans un système dynamique et la seconde présente l'étude des mécanismes de rappel dans les centres d'appels.

Si nous disposons de données détaillées de l'état du système à l'arrivée de chaque client et son temps d'attente réellement observé, nous pouvons alors apprendre une fonction de prédiction des temps d'attente pour ce système en utilisant des algorithmes d'apprentissage machine. Le chapitre 4 présente des prédicteurs de délais d'attente pour les centres d'appels multi-compétences avec des méthodes d'apprentissage machine. Ces prédicteurs sont basés sur une approche heuristique qui combine l'approximation de fonctions, l'apprentissage machine, la simulation, et des idées des prédicteurs QL pour la file unique. Dans ce chapitre, publié en partie dans Thiongane et al. (2015), nous présentons les différents prédicteurs proposés et

étudions la robustesse des prédicteurs face à la variation des taux d'arrivée et face à la variation du staffing des groupes. À la fin de ce chapitre, nous avons comparé nos prédicteurs à un autre prédicteur qui utilise l'apprentissage machine, qui est développé pour les services d'urgence, appelé Q-Lasso (Ang et al., 2016).

Les prédicteurs qui utilisent l'apprentissage machine performent bien dans les systèmes multi-compétences, mais un inconvénient est qu'ils ont beaucoup de paramètres qui doivent être *appris* à l'avance. Cette phase d'entraînement du modèle nécessite une grande quantité de données et du temps de calcul. Ces prédicteurs sont également complexes à implémenter dans la pratique. Au chapitre 5, nous proposons deux prédicteurs DH (E-LES, AvgC-LES), publiés dans Thiongane et al. (2016), simples à implémenter en pratique, et qui possèdent très peu de paramètres . Dans ce chapitre, nous avons présenté en détail les nouveaux prédicteurs et comparé leurs performances avec celles des autres prédicteurs DH et celles des prédicteurs qui utilisent l'apprentissage machine présentés au chapitre 4.

Au chapitre 6, nous proposons de nouvelles idées pour adapter les prédicteurs QL dans les centres d'appels multi-compétences. Nous faisons l'hypothèse que le centre d'appels multi-compétences peut être modélisé par un système alternatif constitué de K modèles de file d'attente indépendants où K est le nombre de type d'appels du centre d'appels multi-compétences. Pour chaque type d'appel, nous avons un groupe d'agent qui traite les appels. La principale difficulté dans ce cas est de déterminer le nombre d'agents de chaque groupe pour avoir l'équivalence des deux modèles. Nous avons proposé plusieurs méthodes pour déterminer le nombre d'agents de chaque groupe. Les erreurs de prédictions obtenues avec ces prédicteurs pour un petit modèle N sans abandons sont plus petites que celles des prédicteurs LES, E-LES et AvgC-LES. Les tests de robustesse des nouveaux prédicteurs face à la variation des taux d'arrivées sont satisfaisants dans ce petit exemple.

Au chapitre 7, nous présentons notre contribution dans l'article d'Ibrahim et al. (2016b) qui propose une modélisation des durées de services dans les centres d'appels. Au début de ce chapitre, nous présentons une revue de la littérature sur la modélisation des durées de service. Par la suite, nous présentations en détails les

nouveaux modèles de durées de service, et la qualité de leurs ajustements sur des données du centre d'appels d'Hydro-Québec. En fin de ce chapitre, nous examinons l'impact opérationnel des modèles par simulation.

CHAPITRE 2

DESCRIPTION D'UN CENTRE D'APPELS ET MESURES DE PERFORMANCES

Dans ce chapitre, nous allons donner une description détaillée des centres d'appels étudiés à la section 2.1. Toutes les mesures de performances, qui sont utilisées dans cette thèse, sont définies à la section 2.2.

2.1 Description du modèle de centres d'appels

Nous considérons des modèles de centres d'appels multi-compétences pour lesquels il y a uniquement des appels entrants. Chaque appel est classé dans l'un des K types d'appels possibles. La classification des appels est en général effectuée par un système automatique appelé "Interactive Voice Response" (IVR). Les agents sont divisés en G groupes. Un agent du groupe $g \in \{1, \ldots, G\}$ a un ensemble de compétences $\mathcal{S}_g \subseteq \{1, \ldots, K\}$ qui définit l'ensemble des types d'appels que cet agent peut servir. Les heures d'ouverture du centre d'appels sont divisées en P périodes d'une durée constante. Par exemple, si le centre d'appels est ouvert de 8:00 à 20:00 et les périodes sont de 30 minutes, nous avons $P = 24$. Nous supposons que les arrivées sont des processus stochastiques ou doublement stochastiques et que nous avons un processus pour chaque type d'appel k. Pour chaque type d'appel k et à la période p, nous considérons que le processus d'arrivée a un taux constant $\lambda_{k,p}$. Dans les exemples étudiés, nous avons considéré des processus d'arrivées Poisson ou Poisson-gamma. Dans le cas des processus de Poisson, le taux est constant fixé sur chaque période p. Dans le cas des processus Poisson-gamma, le taux d'arrivée est une variable aléatoire Gamma sur chaque période p. Le vecteur des taux d'arrivée sur toutes les périodes P est $\lambda_k = (\lambda_{k,1}, \cdots, \lambda_{k,P})$. Ces processus d'arrivée sont supposés indépendants pour tous les types d'appels. Chaque groupe g a un staffing constant $s_{g,p}$ sur chaque période p et $s_g = (s_{g,1}, \ldots, s_{g,P})$ représente le vecteur de

staffing du groupe g sur l'ensemble des P périodes.

Dans les exemples étudiés, les temps de service sont exponentiels de moyenne μ^{-1} ou de loi log-normale de paramètres d'échelle κ et de forme σ. Nous utilisons aussi souvent la moyenne m et la variance v pour caractériser les paramètres de la distribution log-normale. Pour certains des centres d'appels étudiés, nous avons supposé que les agents du groupe sont tous identiques. Dans ce cas-ci, nous définissons une seule distribution du temps de service pour chaque groupe d'agents. Par contre, pour d'autres centres d'appels étudiés, par exemple ceux du chapitre 6 qui est concentré sur la modélisation des durées de service, nous avons supposé que les agents du groupe sont différents. Ainsi, nous définissons une distribution du temps de service pour chaque agent et chaque type d'appels dont il possède la compétence pour le servir. Cette dernière supposition est plus réaliste que la première, car dans la vie réelle, les agents sont des humains. Plusieurs facteurs peuvent influencer leurs performances. En général, les agents qui ont traité de nombreux appels au cours de l'année sont beaucoup plus rapides en moyenne que ceux qui ont manipulé quelques appels.

Pour chaque type d'appel k, les temps de patience sont exponentiels de moyenne ν_k^{-1}. Un client quitte la file d'attente dès que son temps d'attente dépasse son temps de patience. Nous ne modélisons pas les rappels après abandons bien que les abandons dans les centres d'appels peuvent augmenter les futurs taux d'arrivées. Nous supposons que chaque client a besoin uniquement d'un seul type de service et il n'y a pas de possibilité qu'un agent interrompe un appel en service. Il y a une file d'attente par type d'appel. Un nouvel appel de type k est placé à la fin de la file d'attente k, si, à son arrivée, il n'y a pas un agent libre ayant la compétence pour le servir. Les appels de même type sont toujours traités *premier arrivé, premier-servi*. Le routeur attribue les appels aux agents libres selon la politique de routage définie. La politique qui est souvent utilisée est la *politique de routage par priorité*. Selon cette dernière, chaque groupe d'agents définit un ordre de sélection des types appels dont il possède la compétence, et chaque type d'appel a une liste de priorité qui définit l'ordre de sélection des agents. Si plusieurs agents du même groupe sont

disponibles pour traiter un appel, le routeur sélectionne l'agent qui a la plus longue période d'inactivité ; voir Chan et al. (2014) pour plus de détails sur cette politique et sur les politiques de routage en général.

Dans le cas des études de prédiction de délai d'attente, nous utilisons un prédicteur de délai pour chaque type d'appel k. Si un appel de type k doit attendre à la file d'attente, son temps d'attente est immédiatement estimé en utilisant le prédicteur associé à ce type. Le temps d'attente estimé est soit une moyenne conditionnelle à l'état du système, ou bien un temps d'attente déjà observé par un client, ou bien une moyenne conditionnelle des attentes déjà observées par plusieurs clients. Nous ne ré-estimons pas le temps d'attente résiduel du client régulièrement pour mettre à jour la prévision. Nous ne faisons pas une annonce du délai d'attente estimé aux clients et n'étudions pas l'impact de telles annonces dans cette thèse. Les centres d'appels étudiés n'offrent pas une option de rappel aux clients même si les temps d'attente estimés sont longs ou s'il y a congestion du système. Nous supposons aussi que les prédictions de délai n'ont aucune influence sur le comportement des clients ou sur les opérations du centre d'appels. Notre seul but avec ses prédictions est de mesurer l'efficacité des prédicteurs après une simulation du centre d'appels.

Dans nos exemples numériques, nous utilisons souvent trois parmi les modèles canoniques de centres d'appels multi-compétences (Garnett et Mandelbaum, 2000). Le premier est le "modèle V" avec deux types d'appels et un groupe d'agent qui traite les deux types d'appels. Le second est le "modèle N" avec deux types d'appels et deux groupes d'agents, où les groupes ont les ensembles de compétences $\mathcal{S}_1 = \{1\}$ et $\mathcal{S}_2 = \{1, 2\}$. Le groupe 1 peut servir uniquement les appels de type 1 et le groupe 2 peut servir tous les appels. Le troisième est le "modèle W" avec trois types d'appels et deux groupes d'agents. Les groupes ont les ensembles de compétences suivants $\mathcal{S}_1 = \{1, 2\}$, $\mathcal{S}_2 = \{2, 3\}$. Le groupe traite les appels 1 et 2, et le groupe 2 traite les appels de type 2 et de type 3. Les trois modèles sont illustrés à la figure 2.1.

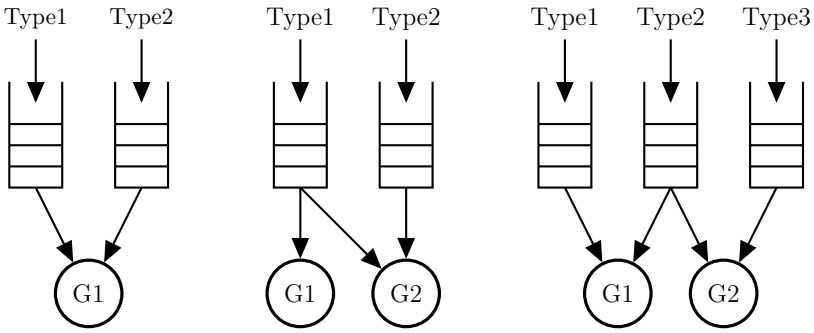

Figure 2.1 : Les modèles multi-compétences V, N et W

2.2 Les mesures de performances

Dans cette section, nous allons définir les mesures de performance qui sont utilisées dans cette thèse. Certains prédicteurs de délai proposés dans cette thèse sont optimisés pour minimiser l'erreur quadratique moyenne ou "Mean Squared Error" (MSE) des prédictions de la journée. Le plus souvent nous utilisons sa version normalisée qui est la racine relative du MSE appelé aussi "Root Relative Mean Squared Error" (RRMSE) pour mesurer les erreurs de prédictions des prédicteurs. Dans nos simulations, nous ne pouvons pas calculer la valeur exacte du MSE mais calculons en général sa valeur empirique, notée ASE et sa racine normalisée appelée le RRASE. Pour quantifier les performances des centres d'appels, nous utilisons aussi le niveau de service SL et le temps d'attente moyen AWT, la probabilité de délai PD, la probabilité d'abandon PA et la longueur moyenne de la file d'attente AQS.

Soit E le temps d'attente prédit d'un client "aléatoire" de type k qui opte pour attendre, choisi au hasard parmi tous les clients sur une infinité de jours dans le modèle, et soit W son temps d'attente réalisé (nous ne considérons pas les clients qui ont abandonné), le MSE pour le type d'appels k est défini comme

$$\mathrm{MSE}_k = \mathbb{E}[(W - E)^2]. \tag{2.1}$$

Puisque nous ne pouvons pas calculer exactement le MSE, nous l'estimons par

sa contrepartie empirique (un estimateur consistant) appelée le *average squared error* (ASE). Nous utilisons la simulation pour calculer le ASE. Soit C le nombre de clients servis de type k qui ont eu à attendre à la file d'attente. Nous notons leurs temps d'attente prédits et réalisés par $E_1, \ldots, E_C$ et $W_1, \ldots, W_C$, respectivement. Le ASE pour le type k est défini comme suit

$$\mathrm{ASE}_k = \frac{1}{C}\sum_{c=1}^{C}(W_c - E_c)^2. \tag{2.2}$$

Notez que nous ne considérons que les clients qui ont vécu un temps d'attente positif, $W_c > 0$, et qui ont attendu jusqu'à ce qu'ils reçoivent un service. Dans certains travaux, la définition du ASE qui inclut également les délais virtuels pour les clients qui ont abandonné, voire par exemple Ibrahim et Whitt (2009a). Il est raisonnable de penser que le ASE va converger vers le MSE avec un grand nombre de réplications des simulations.

Lorsque nous comparons la précision des prédicteurs dans nos expériences numériques, nous générons un ensemble d'observations distinct et indépendant par simulation. Au lieu du ASE, nous utilisons souvent sa racine appelé *"root average squared errors"* (RASE)

$$\mathrm{RASE}_k = \sqrt{\mathrm{ASE}_k}. \tag{2.3}$$

ou bien une version normalisée du ASE, appelé le *"root relative average squared errors"* (RRASE), mesuré sur ce nouvel ensemble de données. Le RRASE pour le type k est

$$\mathrm{RRASE}_k = \frac{\sqrt{\mathrm{ASE}_k}}{(1/C)\sum_{c=1}^{C} W_c}. \tag{2.4}$$

Les RRASE_k des prédicteurs sont mesurées sur le même ensemble de données, de sorte que les W_c et C sont identiques à travers les prédicteurs.

Pour mesurer l'efficacité des prédicteurs par rapport au temps d'attente réellement observé, nous utilisons aussi la moyenne absolue du pourcentage d'erreur, appelée *"mean absolute percentage error"* (MAPE). Il est donné par la formule

suivante :

$$\text{MAPE}_k = 100 \cdot \frac{1}{C} \sum_{c=1}^{C} \left| \frac{W_c - E_c}{W_c} \right|. \tag{2.5}$$

Une mesure de performance que nous observons souvent dans les centres d'appels est le niveau de service ou "service level" (SL) qui est défini par la proportion des appels servis après avoir attendu au plus t unités de temps, où t est une constante appelée le temps d'attente acceptable ou "*acceptable waiting time*" (AWT) (Chan, 2006, Gans et al., 2003). Par exemple, si une entreprise a pour objectif de répondre à 80% de tous les appels dans les 20 secondes qui suivent leur arrivée, alors le seuil de niveau de service est $t = 20$ secondes. Le niveau de service, de même que les autres mesures de performances sont généralement mesurées pour chaque période du centre d'appels (intervalle de temps d'une demi-heure par exemple) et souvent rapportées sur une base quotidienne, hebdomadaire ou mensuelle. Soient $S_B(t)$ le nombre d'appels servis dans un délai inférieur ou égal à t et S le nombre total d'appels servis. Soient N le nombre total d'appels, A le nombre d'abandons et $A_B(t)$ le nombre d'abandons avec un temps de patience inférieur ou égal à t. Il existe plusieurs définitions du SL. Une qui est souvent utilisée est la suivante :

$$h_{\text{SL}} = \frac{\mathbb{E}[S_B(t)]}{\mathbb{E}[N]}. \tag{2.6}$$

Une autre définition du SL qui est aussi souvent utilisée en pratique, exclut $A_B(t)$, les appels qui ont un temps de patience inférieur ou égal au seuil t, du nombre total d'appels à considérer au dénominateur de la formule précédente. Il s'agit de la formule considérée dans cette thèse et choisie par Bell Canada et le Conseil de la radiodiffusion et des télécommunications canadiennes (CRTC) ; voir CRTC (2000). Il est défini comme suit :

$$h_{\text{SL}} = \frac{\mathbb{E}[S_B(t)]}{[N - A_B(t)]}. \tag{2.7}$$

Une mesure de performance que nous observons souvent est le temps d'attente

moyen, appelé “Average waiting time” (AWT), qui est défini par

$$h_w = \frac{\mathbb{E}[W]}{\mathbb{E}[N]}, \tag{2.8}$$

où W est la somme des temps d'attente de tous appels arrivés. Dans cette somme, nous considérons les temps d'attente des appels servis et ceux des appels abandonnés.

La taille moyenne de la file AQS est une autre mesure de performance connexe au temps d'attente. Supposons que le centre ouvre au temps t_i et ferme au temps t_f. AQS est définie comme

$$h_q = \frac{\mathbb{E}[Q]}{\mathbb{E}[t_f - t_i]}, \tag{2.9}$$

où

$$Q = \int_{t_i}^{t_f} b(t)dt$$

et $b(t)$ est la taille de la file d'attente au temps t.

Soit L le nombre total d'appels qui ont attendu à la file d'attente avant d'être servis, la proportion de délais, PD, est alors définie par

$$P_d = \frac{\mathbb{E}[L]}{\mathbb{E}[N]}. \tag{2.10}$$

La proportion d'abandons, PA, est aussi souvent utilisée pour mesurer la performance des centres d'appels. Elle est définie par

$$P_a = \frac{\mathbb{E}[A]}{\mathbb{E}[N]}. \tag{2.11}$$

CHAPITRE 3

REVUE DE LA LITTÉRATURE SUR LA PRÉVISION DES DÉLAIS

Dans ce chapitre, nous allons faire une revue de la littérature sur la prédiction de délai dans les systèmes de service et des travaux qui analysent les effets des annonces de délais d'attente sur le comportement des clients et sur les performances du système. Nous présentons en premier les travaux effectués sur les effets des annonces et nous terminerons par la présentation des travaux faits sur la prédiction du délai dans les systèmes de service en général et dans les centres d'appels en particulier.

3.1 L'effet de l'annonce du délai d'attente

Dans cette section, nous allons examiner les travaux existants sur les effets de l'annonce du délai d'attente dans les centres d'appels téléphoniques. En général, sur ce sujet, on examine les effets de l'annonce des temps d'attente estimés sur le comportement des clients et sur les performances du système. En effet, ces travaux modélisent explicitement la réaction des clients aux annonces de délais d'attente estimés. Le but de cette revue de la littérature est de motiver l'importance de nos recherches sur la prédiction de délai. Ces études ne sont pas faites exactement dans les mêmes modèles de travail que les nôtres (nous travaillons sur des modèles plus complexes), mais les conclusions tirées de ces travaux sont importantes à connaître et peuvent s'étendre dans notre contexte. Les objectifs de ces travaux, sur les annonces de délais pour les centres d'appels, sont faits pour maximiser le taux de satisfaction des clients, améliorer les niveaux de services, diminuer les taux d'abandon, et augmenter le profit des entreprises.

Notons d'abord qu'il existe un ensemble d'articles concentrés sur les effets de l'annonce du délai d'attente dans les systèmes de file d'attente dans d'autres domaines, bien avant leur étude pour les centres d'appels téléphoniques.

De nombreuses études statistiques publiées ont montré l'impact psychologique

négatif des délais d'attente sur les clients dans différents systèmes de services (Dobson et Pinker, 2006, Dubé-Rioux et al., 1989, Hassin, 1986, Taylor, 1994b). Les clients éprouvent souvent de la colère ou du stress à cause de l'incertitude sur le délai d'attente, ce qui peut conduire à diminuer la satisfaction des clients. Des enquêtes ont montré aussi que les clients non informés ont tendance à surestimer leur temps d'attente (Katz et al., 1991). Ces résultats motivent l'utilisation de l'annonce du délai pour diminuer l'impact négatif des attentes. Hui et Tse (1996b) ont observé que fournir des informations de délai aux clients était plus efficace pour les files d'attente avec des durées d'attende "intermédiaires". Lorsque la file d'attente est petite, le temps d'attente est court, alors l'annonce de délai a un impact négligeable. À l'autre extrême, lorsque la file d'attente est longue, annoncer un grand délai ne semble pas apporter une amélioration significative sur la satisfaction des clients. Il est également important de fournir des annonces de délai précis. Mowen et al. (1993) ont observé que les clients ressentent un plus haut niveau d'insatisfaction lorsque leur attente réelle dépasse leur temps d'attente annoncé. D'autre part, les annonces de délais qui sont trop pessimistes peuvent conduire à un plus grand nombre d'abandons. Dans la suite, nous allons parler de l'effet de ces annonces dans les centres d'appels téléphoniques.

Guo et Zipkin (2007) considèrent un centre d'appels modélisé par une file d'attente Markovienne avec un seul serveur (file d'attente M/M/1, les arrivées au centre se font selon un processus de Poisson, les temps de service suivent une distribution exponentielle et on a un seul agent qui répond aux appels). Dans leur modèle, le client qui arrive au centre peut choisir entre deux options. Le premier est de quitter immédiatement le système et le second est de rester dans le système en attendant d'être servi. Les auteurs étudient l'impact de plusieurs niveaux d'information fournis aux clients sur les performances du système. Ils considèrent trois niveaux d'information : (a) aucune information (sur la distribution du temps d'attente), (b) une information partielle (la longueur de la file d'attente), ou (c) information complète (délai d'attente exact). Ils concluent que sous certaines conditions fournir plus d'information sur le délai d'attente peut aider le fournisseur de services à augmenter le

niveau de service ou augmenter la satisfaction des clients. Cependant, dans certains cas, plus d'informations peuvent effectivement nuire à l'un ou à l'autre.

Jouini et al. (2011b) ont étudié l'impact de la précision des annonces sur un système plus complexe et plus réaliste que celui de Guo et Zipkin (2007). Ils utilisent un système multi serveur avec possibilité d'abandon modélisé par une file M/M/s+M. Les arrivées se font selon un processus de Poisson, les temps de service suivent une distribution exponentielle et les abandons se font selon une distribution exponentielle. Les auteurs ont observé l'impact de la précision des annonces sur le taux d'abandon immédiat des clients et le taux d'abandon après avoir séjourné un certain temps dans le système. Comme dans Guo et Zipkin (2007), des annonces de délais avec différents niveaux de précision sont données aux clients à leur arrivée. Après l'annonce de l'information aux clients, ces derniers peuvent ainsi prendre la décision ou bien de quitter immédiatement le système, ou bien d'abandonner après avoir séjourné un certain délai, ou bien d'attendre jusqu'à recevoir le service. Les auteurs ont montré que le taux d'abandons immédiats et le taux des abandons après avoir séjourné un certain temps dans le système sont fonction de la précision du délai d'attente annoncé et de la sensibilité de la réaction des clients face aux délais annoncés. Ils ont observé aussi qu'une plus grande précision dans les annonces n'est pas toujours mieux pour l'amélioration des performances. Dans ce travail, les auteurs ont mené des études analytiques et numériques pour déterminer ce qui devrait être la précision optimale des annonces (la précision qui minimise le nombre d'abandons dans le système). Les conclusions des auteurs dans cet article sont aussi similaires à celles de Mowen et al. (1993) qui sont faites dans un système de service autre qu'un centre d'appels.

Whitt (1999a) étudie un système de file d'attente avec une capacité finie. Le centre d'appels est modélisé par une file M/M/s/r (un système de file d'attente multi serveurs dont la file a une capacité r ; les arrivées se font selon un processus de Poisson et les temps de service suivent une distribution exponentielle). Il compare les performances entre deux modèles. Dans le premier modèle ("modèle 1"), aucune information n'est fournie aux clients à leurs arrivées. Les clients ont la possibilité

d'abandonner immédiatement, ou bien d'abandonner après une certaine attente dans le système (au cas où le temps d'attente réel du client dépasse son temps de patience). Dans le second modèle ("modèle 2"), des délais d'attente estimés en se basant sur l'état du système sont fournis aux clients à leur arrivée au centre d'appels. Il suppose que les délais d'attente estimés et annoncés aux clients sont exacts. Les clients peuvent ou bien abandonner immédiatement ou bien rester dans la file et attendre leur tour pour être servis. Dans le "modèle 2", tous les abandons après une certaine durée dans le "modèle 1" sont remplacés par des abandons immédiats. Ce choix est justifié par le fait que le client connaît son temps de patience, et considère le délai d'attente annoncé comme étant une estimation exacte. Ainsi dès l'annonce du délai d'attente, le client prend la décision de quitter immédiatement quand le délai d'attente annoncé est supérieur à son temps de patience, ou attend son tour d'être servi dans le cas contraire. L'auteur a montré dans l'article que si les deux systèmes ont les mêmes paramètres, le nombre moyen de clients dans le système est plus grand avec le "modèle 1" qu'avec le "modèle 2". Les abandons immédiats des clients à l'arrivée sont plus importants dans le "modèle 2", permettant ainsi de diminuer la congestion du système. Ainsi les clients sont plus susceptibles d'être mis en attente dans le "modèle 1" et plus susceptibles d'être traités sans attente dans le "modèle 2". Avec des exemples numériques, Whitt a aussi évalué certaines mesures de performance (la probabilité qu'un client soit servi, la probabilité de délai, la probabilité d'abandon immédiat et abandon) du système avec les deux modèles, en utilisant un nombre de serveurs s très grand. Ils ont observé que les performances sont très similaires dans les deux systèmes. La seule différence est que dans le modèle avec abandons immédiats, les clients qui ne sont pas servis n'ont pas beaucoup à attendre alors que dans l'autre modèle leurs temps d'attente sont beaucoup plus grands. Whitt (1999a) fournit un support théorique pour l'utilisation des annonces de délai d'attente comme un mécanisme de contrôle pour les fournisseurs de services.

Armony et al. (2009) étudient l'impact des annonces du délai d'attente dans un centre d'appels décrit par un système de file d'attente multi serveurs à tra-

fic intense avec possibilité d'abandon. Ils enlèvent la supposition non réaliste de Whitt (1999a) qui consiste à écarter la possibilité d'abandon après être resté un certain temps dans la file d'attente. Dans leur système, les clients ont la possibilité d'abandonner immédiatement ou bien d'abandonner après être restés une certaine durée dans le système. Les auteurs utilisent les approximations dans les modèles fluides déterministes développés par Whitt (2006) pour déterminer les performances approximatives du système. Dans un premier temps, ils déterminent le temps d'attente moyen à l'état d'équilibre (FD) et l'utilisent pour faire des annonces de délai d'attente aux clients. Dans un second cas, ils annoncent le temps d'attente réel du dernier client à entrer en service (DLS) comme le temps d'attente estimé pour tout nouvel appel. La comparaison des deux modèles a permis de conclure que l'annonce du délai d'attente dépendant de l'état du système (DLS) est plus précise que l'annonce d'un délai d'attente fixe (FD) pour l'ensemble des clients. Le nombre d'abandons est plus grand dans le premier cas que dans le second. De même, l'erreur quadratique moyenne (MSE) des prédictions est plus grande dans le premier cas que dans le deuxième.

Jouini et al. (2011a) étudient l'impact de la précision des annonces sur le taux d'abandon immédiat des clients et le taux d'abandon après avoir séjourné un certain temps dans le système, comme dans Jouini et al. (2011b), à la différence qu'ici les clients peuvent réagir aux annonces par une modification de leur temps de patience. Ils utilisent un centre d'appels avec des clients impatients. Les clients réagissent par des abandons immédiats et par des abandons après une certaine attente, particulièrement quand ils réalisent que le délai d'attente réel excède le délai qui leur est initialement annoncé. Dans cet article, les auteurs étudient deux modèles. Dans le premier modèle noté "modèle 1", aucune annonce du délai d'attente n'est fournie aux clients. Dans le deuxième modèle, "modèle 2", des informations sur le délai d'attente estimé sont annoncées aux clients à leurs arrivées. Du fait de l'impossibilité de trouver une estimation de délai exacte, un délai correspondant à une probabilité de couverture β est communiqué au client. Le délai d'attente E_i estimé et communiqué au client i satisfait la contrainte $\mathbb{P}[W_i \leq E_i] \geq \beta$ où W_i est le vrai

temps d'attente (temps d'attente réel) du client i dans le système. Dans le "modèle 2" analysé par Whitt (1999a), une fois que le nouveau client accepte de rejoindre la file, il n'abandonne plus jamais. Whitt avait remplacé tous les abandons après une certaine attente du "modèle 1" par des abandons immédiats dans le "modèle 2" et ceci était justifié si on suppose que les délais d'attente estimés sont exacts. C'est-à-dire que l'estimateur est parfait. Mais comme nous utilisons des processus stochastiques, la prédiction et le délai annoncé ne sont pas exacts. Puisque les clients connaissent cela, nous pouvons envisager que beaucoup d'autres choses se passent. Ici les auteurs autorisent le client (indexé par i) à mettre à jour son temps de patience t_i en réponse au délai d'attente annoncé E_i. Le temps de patience est modélisé par la valeur suivante : $\theta t_i + (1-\theta)E_i$ pour $\theta \geq 0$. Pour un client i qui ne met pas à jour son seuil de patience t_i après l'annonce du délai alors $\theta = 1$ ("no update case"). Une autre possibilité est que le client met complètement à jour son seuil de patience et remplace t_i par le délai annoncé, représenté par $\theta = 0$ ("update case"). Le cas où $\theta > 1$ peut correspondre à un cas où les annonces entraînent une augmentation du temps de patience des clients. Dans leur modèle, ils n'autorisent pas $\theta < 0$. Les auteurs ont montré que l'annonce d'un délai d'attente avec une couverture élevée est importante quand la réaction des clients au délai est assez élevée ou lorsque la prévention des abandons de clients est jugée essentielle (contrainte de niveau de service stricte), ou lorsque la congestion du système est élevée. Par une étude analytique, ils ont montré qu'un compromis entre les abandons après une certaine attente et les abandons immédiats peut être réalisé en choisissant la couverture d'annonce. Ils ont montré aussi qu'une couverture élevée n'est pas nécessairement meilleure pour les fournisseurs de services et que la couverture des annonces doit être soigneusement contrôlée en présence de différentes réactions des clients. Ce contrôle supplémentaire, s'il est correctement utilisé, fournit aux gestionnaires un moyen d'améliorer les performances, en particulier si la réaction des clients aux annonces est forte et si le système est petit ou surchargé.

Jouini et al. (2011a) étudient un système où le délai d'attente estimé n'est pas fourni immédiatement aux clients à leur arrivée, mais après une courte période

(passée soit en attente ou occupé par le système). Les auteurs étudient l'impact de ce report sur la capacité du gestionnaire à influencer le comportement des clients en leur communiquant des informations de congestion non vérifiables. Ils étudient aussi l'impact de ce report sur les profits de l'entreprise et les utilités des clients. Ils considèrent un système où les clients qui arrivent sont servis en deux étapes. La première étape, qui n'est pas nécessairement pour le service, ne nécessite aucune ressource humaine, est généralement effectuée par un serveur automatique ou "*Interactive Voice Response*" (IVR) dans le cadre des centres d'appels. Cette étape est habituellement utilisée pour fournir aux clients des informations générales ainsi que pour recueillir des informations auprès du client d'une manière efficace. Le délai de traitement avec l'IVR est considéré comme le mécanisme qui fournit le délai du report et est utilisé pour étudier l'impact du report de l'annonce sur les délais d'attente. La deuxième étape, qui génère de la valeur pour l'entreprise et le client, exige des ressources humaines. Après que les clients ont terminé leurs interactions avec l'IVR, le gestionnaire leur fournit un message indiquant la durée de l'attente dans le système. À ce stade, le client peut décider de rejoindre ou d'abandonner le système. S'il décide de rejoindre, il entre dans la file d'attente multi serveur pour être servi et n'a plus la possibilité d'abandonner. Ils ont d'abord montré que si l'entreprise a le contrôle total (c'est-à-dire peut demander à un client d'abandonner immédiatement ou bien de rejoindre la file pour être servi par un agent), sous certaines conditions, une politique de contrôle d'admission optimale peut être obtenue et ce report peut aider l'entreprise à améliorer son profit. Cependant, en pratique, il est difficile et aussi très coûteux de demander à un client de quitter une fois admis dans le système. Par la suite, ils ont montré aussi que ce délai peut réellement aider l'entreprise à créer de la crédibilité et entraîner un équilibre (en utilisant des niveaux d'information supplémentaire non précisée). Toutefois, ce délai peut également nuire à l'équilibre du système et à la crédibilité du gestionnaire si l'entreprise est plus sophistiquée dans ses stratégies.

Hui et Tse (1996b) ont observé que lorsque la file d'attente est longue, annoncer un grand délai ne semble pas apporter une amélioration significative sur la satis-

faction des clients et peut même dans certains cas entraîner une augmentation des abandons dans le système. Dans une telle situation, il est préférable d'offrir aux clients la possibilité d'être rappelé plus tard. Des enquêtes effectuées sur plusieurs centres de contacts (Advice, 2014) ont montré que plus de 60% des clients préfèrent être rappelés que de rester en attente pendant plusieurs minutes. En plus d'augmenter le taux de satisfaction, et d'éviter les longues attentes pour les clients, une bonne stratégie de rappels (une bonne politique de routage des appels en attente à la file et des rappels) peut aider à équilibrer les charges du système, diminuer les abandons, éviter les répétitions d'appels, et augmenter le niveau de service des systèmes. Actuellement, beaucoup de centres d'appels utilisent les délais estimés pour faire des annonces et proposer le rappel si nécessaire. La plupart des logiciels pour la gestion de centre d'appels (Five9, Virtual Contact Center, VanillaSoft, etc) supportent l'option de rappel.

Nous allons maintenant examiner des travaux qui combinent les annonces de délai d'attente avec la proposition d'une option de rappel. Dans ces travaux, en plus de l'annonce du délai, une option de quitter et d'être rappelé plus tard est proposée au client si le temps d'attente estimé est supérieur à un certain seuil de S unités de temps. Ce seuil S (par exemple 30 minutes), qui est synonyme d'une longue attente ou d'une congestion du système, est fixé par le gestionnaire. Avec un délai d'attente estimé $D \geq S$, la probabilité d'abandonner avant de recevoir du service peut être (dépendant de S) largement supérieure à la probabilité d'attendre à la file jusqu'à être servi. Dans le cas où le client choisit d'être rappelé, différents mécanismes de rappel peuvent être utilisés. Dans certains cas, le rang à la file d'attente du client qui opte pour le rappel est toujours maintenu. Autrement dit le client quitte réellement la file d'attente, mais son rang à la file est toujours virtuellement maintenu. Ainsi, le rappel du client est effectué quand son tour arrive à la file. Ce mécanisme est par exemple utilisé au centre d'appel Hydro-Québec et au centre d'appel de la compagnie de téléphone FIDO. Dans d'autres cas, le rappel du client est effectué après une certaine durée. Cette durée peut-être une valeur fixe (elle reste toujours la même pour tous les clients de rappel) ou bien une variable

dépendant de l'état du système (par exemple quand le système devient vide). Les clients qui sont en attente dans la file sont souvent appelés les "clients réels", et ceux qui acceptent d'être rappelés sont appelés "les clients virtuels". Beaucoup de travaux se sont penchés sur la recherche de politiques de routage "*optimales*" des différents types d'appels (réels et virtuels) qui minimisent la durée moyenne des attentes pour les "clients réels", et qui maximisent le niveau de service pour les "*clients virtuels*" dans des systèmes de file d'attente avec option de rappel. Dans cette section, nous présentons quelques travaux qui utilisent les délais d'attente estimés pour déterminer la politique de routage optimale des différents types de clients. Cependant, il faut noter qu'il existe des travaux qui n'utilisent pas les temps d'attente pour déterminer la politique de routage optimale dans des systèmes avec des clients réels et virtuels ; voir par exemple Dudin et al. (2013), Gans et Zhou (2003), Kim et al. (2012) et Ding (2016).

Armony et Maglaras (2004a) examinent un centre d'appels modélisé par une file d'attente multi serveurs sans abandons qui offre une option de rappel aux clients. Les clients arrivent selon un processus de Poisson, les temps de service sont exponentiels. Il y a s serveurs identiques qui servent les appels. À son arrivée au centre, le client est informé de deux délais d'attente. Le premier est le délai d'attente estimé dans le cas où le client choisit d'attendre dans la file. Le second est le délai limite d pour être rappelé par un agent (un appel sortant) dans le cas où il choisit l'option de rappel. Dans leur modèle, le client peut choisir entre : (i) rejoindre la file d'attente et attendre d'être servi, (ii) laisser un message pour le service de rappel, ou (iii) abandonner immédiatement et ne pas entrer dans le système. La décision du client est prise est à l'issue du calcul de l'utilité associé à chaque choix et il choisit toujours celle qui a la plus grande utilité. Les auteurs considèrent un modèle avec deux classes de clients avec chacune sa file d'attente. Les clients qui optent pour le service en temps réel constituent la *classe 1* et ceux qui optent pour l'option de rappel sont de la *classe 2*. Les clients arrivent au centre selon un processus de Poisson de taux λ. On note par $\lambda_1(\mathrm{S})$, $\lambda_2(\mathrm{S})$, $\lambda_0(\mathrm{S})$ les taux d'arrivée, dépendant de l'état du système S, avec lesquelles les clients se joignent à la classe 1, à la classe

2, ou abandonnent, respectivement, avec $\lambda = \lambda_1(S) + \lambda_2(S) + \lambda_0(S)$.

Le système décrit ci-dessus peut être représenté par un modèle V ; un système multi-compétences avec deux types d'appels et un seul groupe d'agents (Gans et al., 2003, Garnett et Mandelbaum, 2000). L'objectif principal des auteurs est de déterminer une politique de routage optimale (c.-à-d. une politique qui maximise le niveau de service et minimise le temps d'attente pour les clients réels, tout en respectant la contrainte du délai limite de démarrage du service pour les clients virtuels) des deux types d'appels (*classe 1* et *classe 2*). Dans un tel système (malgré une simplification non réaliste qui suppose qu'il n'y a pas d'abandon après être resté une certaine durée dans la file d'attente), la tâche d'estimer le délai d'attente pour un client réel conditionnel à l'état du système est assez complexe. Sa dépendance aux taux d'arrivée, à l'état du système et à la structure de la politique de routage proposée rend ce calcul très complexe parce que les délais d'attente de la classe 1 dépendent de futures arrivées de classe 2, et les deux sont des fonctions de l'état qui change avec le temps. Pour simplifier davantage, les auteurs se concentrent sur le cas particulier des grands systèmes (grand s) à trafic intense, qui caractérise certains centres d'appels. Dans un tel modèle, la situation se simplifie considérablement. Cela est dû à l'observation suivante : les grands systèmes multi-serveur bénéficient d'une forme d'économie statistique d'échelle ; en particulier, le temps d'attente des clients à la file réel décroît vers zéro, même si le système est approché à trafic intense. Par ailleurs, l'état du système (nombre de serveurs occupés et le nombre de clients dans la file d'attente) ne change pas de manière significative durant chaque courte période d'attente.

Dans cette thèse, nous proposons des méthodes des prédictions au chapitre 4 et 5 qui sont indépendantes du routage, de la taille du système et du régime pour les systèmes multi-compétences. Armony et Maglaras proposent une politique de routage optimale qui utilise seulement l'information sur la longueur de la file d'attente pour sélectionner le type d'appels à traiter. Elle donne la priorité à la classe 2 quand la longueur de sa file d'attente excède un certain seuil et à la classe 1 autrement. Cette politique de routage est asymptotiquement optimale dans le

sens où elle minimise le temps d'attente pour les clients réels (classe 1) tout en respectant la contrainte de date limite de démarrage des services des clients de rappel (classe 2), (min $\mathbb{E}(W_1)$ sujet à $W_2 \leq D_2$) où W_i le temps d'attente d'un client qui se trouve à la file i le plus longtemps au temps t). Cette politique optimale a été trouvée en utilisant la proposition prouvée par Maglaras et Mieghem (2004) qui stipule : Soit $A_i(t)$ le nombre total de clients qui sont arrivés dans la file i durant la période de temps $[0, t]$, et $Q_i(t)$ la longueur de la file i à l'instant t, alors nous avons ceci :

$$W_2 \leq D_2 \ \forall \ t \Longleftrightarrow Q_2(t) \leq A_2(t) - A_2(t - D_2) \ \forall \ t \tag{3.1}$$

C'est-à-dire aucun client de classe 2 n'a été en attente pendant plus de D_2 unité de temps si et seulement si tous les clients actuellement en attente à la file 2 sont arrivés dans les dernières D_2 unités de temps. Par conséquent, avec $Q_2(t) \leq A_2(t) - A_2(t - D_2)$, le seuil approprié à utiliser est $\theta(t) = A_2(t) - A_2(t - D_2)$ et la politique correspondante est spécifiée comme suit : si $Q_2(t) \geq \theta(t)$, donner la priorité à la classe 2, sinon donner la priorité à la classe 1. Maglaras et Mieghem utilisent une méthode d'estimation du délai d'attente des clients réels qui est asymptotiquement optimale (si le système est dans un état d'équilibre). En supposant avec optimisme que l'état de la file et le taux d'arrivée sont en effet constants au cours du temps qu'un client séjourne dans la file 1, on peut estimer le délai d'attente dépendant de l'état comme suit : Soit Q_1 la longueur de la file d'attente de la classe 1 (son taux d'arrivée est λ_1), une version locale de la loi de Little montre que le temps d'attente de classe 1 peut être estimé par

$$W_1 = \frac{Q_1}{\lambda_1}. \tag{3.2}$$

Ils concluent qu'en informant les clients sur les délais d'attente estimés, les gestionnaires peuvent bien contrôler la congestion, et équilibrer les charges entre les deux classes de clients.

Armony et Maglaras (2004b) ont travaillé sur le même modèle que dans Armony et Maglaras (2004a). Ils considèrent un système à trafic intense dans son état d'équilibre. Dans cet article, l'information donnée aux clients est le temps d'attente moyen à l'équilibre du système, alors que dans Armony et Maglaras (2004a), l'information reçue par le client est une estimation du délai conditionnelle à l'état du système. Ils supposent que les informations fournies aux clients sont exactes, c'est-à-dire que l'erreur de prévision est nulle. En comparant les résultats des deux systèmes, les auteurs ont montré que plus d'information augmente le taux d'utilisation global du système tout en offrant une meilleure qualité de service aux clients. Le même niveau de service est aussi gardé pour les clients qui ont opté pour le rappel.

Résumé des recommandations de ces études.

Plusieurs types de modèles (M/M/1, M/M/s, M/M/s/r, M/M/s+M, Modèle V) ont été étudiés dans cette revue de la littérature sur les annonces de délais aux clients. Les conclusions tirées de ces travaux sont importantes à connaître et montrent qu'il est important d'avoir de bons prédicteurs de délais pour les systèmes de service pour améliorer les performances de ces systèmes et augmenter la satisfaction des clients. Des recommandations sont formulées dans chaque situation pour éviter de détériorer les performances du système : (i) Pour augmenter la satisfaction des clients, et diminuer les abandons dans un système de service, il est toujours préférable de fournir aux clients des informations de la longueur de la file d'attente que de ne fournir aucune information ; (ii) Fournir des informations sur le délai d'attente contribue mieux à améliorer les performances du système que de fournir l'information sur la longueur de la file d'attente ; (iii) Si la variance sur les temps attentes observées par les clients est grande, alors il est préférable de fournir des informations de délai d'attente conditionnelle à l'état du système que d'informer les clients du temps d'attente moyen du système ; (iv) Cependant fournir de fausses informations de délais aux clients peut détériorer considérablement les performances

du système. Il est important d'avoir des prédictions assez précises ; (v) Si les temps d'attente estimés sont longs, l'annonce du délai d'attente ne contribue pas à améliorer les performances du système ni à augmenter la satisfaction des clients. Dans cette situation la proposition d'une option de rappel combinée aux annonces peut beaucoup contribuer à équilibrer les charges du système, diminuer les abandons, améliorer les performances du centre d'appels.

Nous notons d'après ces études et recommandations, qu'il est nécessaire d'avoir de bons prédicteurs de délais pour les systèmes de service. Dans la section suivante, nous allons présenter les travaux sur les méthodes de prédiction dans les systèmes de service et plus particulièrement dans les centres d'appels.

3.2 La prédiction du délai d'attente des clients

La seconde partie de cette revue de littérature met l'accent sur la prédiction du temps d'attente des clients dans les systèmes de service où les prédictions pourraient être utilisées pour faire des annonces de délai d'attente aux clients. Notons d'abord qu'il existe un ensemble d'articles concentrés sur la prédiction du délai d'attente dans les manufactures bien avant leur étude dans les centres d'appels téléphoniques ; voir par exemple Morton et Vepsalainen (1987), Ornek et Collier (1988), Shanthikumar et Sumita (1988). En général, dans ces travaux, on estime la durée de fabrication des produits dans les ateliers en plusieurs étapes. Dans notre contexte, on estime le délai d'attente d'un client avant son entrée en service dans un centre d'appels. Plusieurs travaux ont été réalisés dans ce cadre. Cependant, il faut noter que la plupart de ces travaux sont faits pour des systèmes de service avec un seul type de client (une seule file d'attente). Les travaux qui sont faits pour les systèmes complexes avec plusieurs classes de clients et plusieurs groupes de serveurs (systèmes multi-compétences) sont rares. Le peu qui existe est en général fait pour des systèmes très particuliers. Une contribution majeure de cette thèse est le développement de nouveaux prédicteurs qui peuvent être utilisés dans tous les systèmes multi-compétences. Il faut noter aussi que les prédicteurs, qui

utilisent seulement l'historique du système, développés pour les systèmes avec une seule classe de clients, peuvent s'étendre au cas multi compétence. Mais malheureusement, ces prédicteurs donnent la plupart du temps de mauvaises performances pour les systèmes multi-compétences actuels. La revenue de littérature dans ce contexte sera divisée en deux parties. Dans la première partie, nous allons présenter les travaux pour les systèmes avec un seul un type de client et dans la seconde partie, nous allons présenter les travaux faits pour les systèmes multi-compétences.

3.2.1 Les prédicteurs pour les systèmes avec un seul type de clients et une seule file d'attente

Le prédicteur de délai le plus simple qui peut être utilisé pour prédire le temps d'attente d'un client est celui qui ne regarde aucune information et qui prend le temps d'attente moyen global sur tous les clients. Il est appelé le prédicteur NI ou *"Non-Information predictor"* (Armony et Maglaras, 2004b, Ibrahim et Whitt, 2009a). En général, les performances de ce prédicteur sont mauvaises sauf pour les grands systèmes à trafic intense à l'état d'équilibre pour lesquels les temps d'attente des clients sont très similaires.

Des prédicteurs qui utilisent les informations du système sont étudiés dans Whitt (1999b). Dans cet article, l'auteur se concentre sur l'estimation des délais d'attente dans divers systèmes de file d'attente multi serveur. Whitt suppose que le système reçoit un seul type de clients et que les serveurs sont tous identiques. Il a travaillé sur deux modèles différents. Le premier sans abandon et le second avec abandon. Il a montré qu'on peut bien estimer le délai d'attente W d'un nouveau client ou d'un client déjà dans la file si nous connaissons les informations sur l'état du système. Ces informations sont le nombre de clients en attente de service qui précèdent le client à la file, le nombre de serveurs, et le taux de sortie du système. Par exemple pour le modèle GI/M/s/r (GI indique le processus d'arrivée, les temps de service sont exponentiels de moyenne μ^{-1}, s serveurs, r est la capacité de la file), à chaque fois que tous les serveurs sont occupés, le temps jusqu'à la prochaine fin de service est une exponentielle de moyenne de $1/s\mu$, indépendamment du passé.

Par conséquent, le temps d'attente avant le démarrage du service pour une nouvelle arrivée avec $s+k$ clients dans le système est la somme de $k+1$ variables aléatoires exponentielles i.i.d de moyenne de $1/s\mu$ chacune, qui suit une distribution d'Erlang. Ainsi, l'espérance et la variance du temps d'attente W, conditionnelle aux informations de l'état du système, sont données par :

$$\mathbb{E}[W] = \frac{k+1}{s\mu} \quad \text{et} \quad \text{Var}[W] = \frac{k+1}{(s\mu)^2} \tag{3.3}$$

Ici on pourrait donner au client la loi de probabilité de son temps d'attente. Par exemple un histogramme ou graphique de la densité.

En prenant le même modèle avec possibilité d'abandon et en supposant que le client à la position j de la file peut abandonner avec un taux δ'_j alors le délai d'attente du client ayant trouvé $k+1$ clients dans le système, peut être représenté comme la somme de $k+1$ exponentielles, mais pas identiquement distribuées. Le taux total d'abandon quand il y a k clients dans la file est donné par :

$$\delta_k = \sum_{j=1}^{k} \delta'_j \tag{3.4}$$

et l'espérance et la variance du délai d'attente W sont données par :

$$\mathbb{E}[W] = \sum_{j=1}^{k} \frac{1}{s\mu + \delta_j} \quad \text{et} \quad \text{Var[W]} = \sum_{j=1}^{k} \frac{1}{(s\mu + \delta_j)^2} \tag{3.5}$$

Les prédicteurs qui utilisent la longueur de la file d'attente et les paramètres du système pour estimer le délai d'attente sont souvent appelés "*Queue Length predictor*" (QL).

La supposition que les serveurs sont tous identiques et leur nombre constant dans le temps, qui est faite pour développer les prédicteurs QL, n'est pas toujours réaliste dans les systèmes de service. Par exemple dans un centre d'appels, le nombre de serveurs et les moyennes du temps de service peuvent être variables dans le temps parce que les serveurs sont des êtres humains qui servent dans différentes

périodes et pourraient bien avoir différentes distributions du temps de service. Les durées de service des agents ne sont pas en général exponentielles comme on le suppose souvent. Dans de telles situations, les prédicteurs QL ne sont pas adaptés et des prédicteurs qui ne dépendent pas des paramètres du système peuvent être préférables. Ibrahim et Whitt (2008) ont développé des prédicteurs qui n'utilisent aucun paramètre du système et qui sont très simples à implémenter en pratique. Pour estimer les temps d'attente des clients, les auteurs proposent des prédicteurs qui utilisent les délais d'attente déjà vécus par les anciens clients du système. Les différents prédicteurs considérés dans cet article sont : (i) le délai du dernier client à entrer en service ou "*the delay of the last customer to enter service*" (LES), (ii) le délai enregistré par le client à la tête de la file ou "*the delay experienced so far by the customer at the head of the line*" (HOL), (iii) le délai du dernier arrivé parmi les clients qui ont récemment terminé leur service ou "*the delay experienced by the customer to have arrived most recently among those who have already completed service*" (RCS). Ibrahim et Whitt ont comparé la précision des différents prédicteurs selon le critère du MSE pour le modèle GI/M/s, en insistant sur les grands s. Ils observent que les prédicteurs LES et HOL sont très similaires et sont plus précis que RCS. Leur comparaison avec le prédicteur QL (équation 3.3) montre qu'ils sont légèrement moins précis que ce dernier. Les prédicteurs DH fournissent environ les mêmes performances que QL lorsque le processus d'arrivée à une très faible variabilité. Dans la pratique les prédicteurs DH sont attrayants et ont l'avantage d'être robuste, car ils répondent automatiquement aux changements de paramètres du système. Cependant, nous notons que les performances de ces prédicteurs se dégradent lorsque la variation du processus d'arrivée augmente.

Pour améliorer leurs performances dans les systèmes avec variation importante dans le processus, Ibrahim et al. (2016a) proposent deux ajustements du prédicteur LES. Le premier est un prédicteur LES proportionnel à la longueur de la file d'attente observée (P-LES). Soient Q_{LES} le nombre de clients dans la file d'attente lorsque le client LES arrivé, x le délai d'attente du client LES, et Q le nombre de clients dans la file d'attente en avant du nouveau client arrivé. Pour tenir compte

de la variation de la longueur de la file d'attente, les auteurs considèrent (comme une heuristique) un prédicteur qui multiplie x par le rapport Q/Q_{LES}. Ainsi le délai d'attente W du nouveau client est estimé par

$$\mathbb{E}[W] = x\frac{Q}{Q_{\text{LES}}}. \tag{3.6}$$

Le second prédicteur est proposé pour les modèles fluides avec une seule file d'attente (A-LES). Soient λ, μ^{-1}, et ν^{-1} le taux d'arrivée, la moyenne des temps de service, et la moyenne des temps de patience, respectivement. Le prédicteur A-LES prédit l'espérance du délai W par

$$\mathbb{E}[W] = \frac{1}{\nu}\ln(\rho + 1 - \rho e^{-\nu x}), \tag{3.7}$$

où $\rho = \lambda/(\mu s)$, s est le nombre de serveurs, et x est le délai du prédicteur LES. Notons que ce prédicteur dépend des paramètres du système à cause de ρ et ν. Les études comparatives menées sur ces prédicteurs pour plusieurs modèles M/M/s+M, montrent que LES est plus précis que P-LES, mais moins précis que A-LES. QL est toujours plus précis que P-LES et A-LES.

Un autre prédicteur souvent utilisé en pratique est le prédicteur Avg-LES qui prédit le temps d'un nouveau client par une moyenne de plusieurs LES (Armony et al., 2009, Dong et al., 2016). Il retourne le délai moyen vécu par les N derniers clients qui sont entrés en service, pour un nombre entier $N > 0$ fixe, ou une variable aléatoire qui représente le nombre de clients qui sont entrés en service dans les T dernières unités de temps. Une plus grande valeur de N ou une plus grande fenêtre de temps T augmente le lissage et peut ainsi réduire la variance du prédicteur, mais ce grand décalage le plus souvent conduit à des prédictions moins précises, car il utilise des informations anciennes (moins pertinentes). En particulier, les prédictions sont plus susceptibles d'être basées sur les attentes de clients qui ont vu une file d'attente très différente devant eux quand ils sont arrivés. Les N ou T peuvent être prises comme tous égales, mais il pourrait aussi avoir du sens à prendre

un grand N ou un petit T pour les classes de clients les plus fréquentes. Dans nos expériences, nous avons constaté que le meilleur choix de N était habituellement $N = 1$, qui est équivalent à LES.

Avg-LES peut être généralisé à une *moyenne pondérée* des derniers temps d'attente. Nous choisissons une séquence de poids non négatif $\phi_1, \phi_2, \ldots$, généralement non croissant et qui converge vers 0, et tel que $\sum_{i=1}^{\infty} \phi_i = 1$. Ensuite, nous prédisons le temps d'attente d'un client qui arrive par

$$D = \sum_{i=1}^{\infty} \phi_i W_i, \tag{3.8}$$

où W_i est le temps d'attente du i-ième dernier client qui a commencé le service (le LES pour $i = 1$, le précédent pour $i = 2$, etc.). Ce prédicteur a de nombreux paramètres (les poids) dans sa forme générale, mais ce grand nombre de paramètres peut être facilement réduit en mettant des contraintes sur les poids.

En prenant $\phi_i = 1/N$ pour $i = 1, \ldots, N$ et $\phi_i = 0$ pour $i > N$, nous retrouvons Avg-LES. Si nous prenons $\phi_i = \alpha(1-\alpha)^{i-1}$ à la place, pour un facteur de lissage $\alpha \in (0, 1]$, nous obtenons une *moyenne exponentielle* (ESAvg-LES) au lieu d'une moyenne ordinaire. Pour $\alpha = 1$, nous retrouvons LES. Pour $\alpha < 1$, l'implémentation doit être approximative, parce que dans la pratique, nous avons seulement un nombre fini de délais passés. Dans notre implémentation de ESAvg-LES, nous initialisons un prédicteur S à -1, et nous mettons à jour S comme suit. Chaque fois qu'un nouveau client commence le service après un temps d'attente W, nous mettons S à W si $S = -1$, sinon nous le mettons à jour par

$$S := \alpha W + (1-\alpha)S. \tag{3.9}$$

Quand un client entre dans la file d'attente, son temps d'attente est prédit par le S courant. Si $S = -1$, on retourne la valeur du prédicteur LES. Selon nos expériences, le meilleur choix de α est généralement proche ou égal à 1.

Au chapitre 5 de cette thèse, nous proposons deux prédicteurs DH notés E-

LES et AvgC-LES qui s'adaptent rapidement aux variations dans le système. Nous comparons avec les prédicteurs LES, Avg-LES, P-LES, ESAvg-LES. Les résultats montrent que nos prédicteurs sont largement plus performants que ces prédicteurs DH présentés dans cette revue de la littérature.

Ibrahim et Whitt (2009c) ont développé divers prédicteurs QL pour plusieurs types de modèles avec abandons. Ils considèrent des modèles fluides dans des régimes ED à trafic intense. Pour être plus précis, ils utilisent la limite des régimes ED à trafic intense, selon un modèle asymptotique développé par Halfin et Whitt (1981), pour proposer des prédicteurs de délai pour divers modèles GI/GI/s+GI. Les auteurs ont d'abord proposé le prédicteur Markovien noté QL_m pour le modèle GI/M/s+M. Ce prédicteur est une variante du prédicteur QL proposé par Whitt (1999b) pour les modèles avec abandon. Il tient compte des abandons en supposant que les temps de patience des clients sont des exponentiels i.i.d de taux α. Les temps de service sont des exponentiels i.i.d de moyenne μ^{-1}. Il estime le temps d'attente W du client qui a trouvé k autres clients en attente dans la file par :

$$\mathbb{E}[W_{\mathrm{QL}_m}] = \sum_{j=1}^{k} \frac{1}{s\mu + j \cdot \alpha}. \tag{3.10}$$

Les auteurs ont montré que l'estimateur QL_m est assez précis, selon le critère du MSE, pour le modèle GI/M/s+M, mais il n'est pas précis pour le modèle plus général GI/M/s+GI surtout si la distribution des abandons est loin de la distribution exponentielle. Par la suite, ils ont proposé un autre prédicteur très simple, QL_r qui multiplie l'estimation du QL standard par une constante β dépendant du modèle, basé sur des approximations fluides dans la limite des régimes ED à trafic intense. Le prédicteur estime le temps d'attente du nouveau client par :

$$\mathbb{E}[W_{\mathrm{QL}_r}] = \beta \cdot \frac{k+1}{s\mu}. \tag{3.11}$$

En pratique, il est possible d'apprendre la constante en observant les valeurs réelles des attentes et les valeurs estimées de l'estimateur QL standard pour tous les clients

sur de vraies données. Si de telles données n'existent pas, on peut les générer par une simulation du modèle. Les auteurs ont montré par la simulation que QL_r est assez performant pour les grands systèmes à trafic intense (c'est à dire à chaque fois que les approximations fluides sont appropriées). Lorsque la distribution du temps des abandons est loin de la distribution exponentielle, le prédicteur QL_r fonctionne beaucoup mieux que le prédicteur QL_m.

Ibrahim et Whitt (2009c) ont proposé un nouveau prédicteur, noté QL_{ap}, pour le modèle GI/M/s+M(k). Les temps de patience des clients sont exponentiels indépendants de moyenne α_k^{-1} dépendants de la longueur de la file système k. Ils utilisent les approximations des modèles fluides et prédisent le temps d'attente W d'un nouveau client par

$$\mathbb{E}[W_{\mathrm{QL}_{ap}}(k)] = \sum_{j=1}^{k} \frac{1}{s\mu + \alpha_k - \alpha_{k-j}}. \tag{3.12}$$

Les résultats des simulations ont montré que l'estimateur QL_{ap} est toujours le prédicteur le plus efficace. Il coïncide avec QL_m dans le cadre du modèle GI/M/s+M. Il performe aussi assez bien pour les distributions des temps de patience non exponentiels. Cependant, il est important de noter que les prédicteurs QL_r et QL_{ap} nécessitent une connaissance du taux d'arrivée λ, qui nécessite un certain degré de stationnarité (le taux d'abandon est une fonction qui dépend du taux d'arrivée λ). Ces prédicteurs sont efficaces si le taux d'arrivée effectif ne varie pas trop vite. Enfin, les auteurs ont étudié le prédicteur LES pour les modèles à trafic intense, et ont montré qu'il est très efficace à l'état d'équilibre du système pour tous les modèles considérés. Dans Ibrahim et Whitt (2010), les auteurs étudient les mêmes systèmes que Ibrahim et Whitt (2009c) en supposant que la longueur de la file d'attente n'est pas observable. Pour estimer la longueur de la file d'attente, ils utilisent le délai du prédicteur HOL, le taux d'arrivée et le temps moyen entre deux départs successifs du système.

Ibrahim et Whitt (2011) ont développé de nouveaux prédicteurs pour les systèmes multi serveurs (avec abandons) encore plus complexes que ceux étudiés dans

Ibrahim et Whitt (2010). Ils supposent qu'en plus du taux d'arrivé, le nombre de serveurs varie aussi dans le temps. Nous avons ici un système multi serveur avec un processus d'arrivées non stationnaire et un nombre de serveurs variable dans le temps. Les auteurs ont commencé d'abord par montrer que les prédicteurs existants qui ne tiennent pas compte de la variation du nombre de serveurs sont biaisés pour le modèle M(t)/M/$s(t)$+GI. Ils ont ensuite proposé plusieurs prédicteurs pour ce modèle. Le plus efficace est un prédicteur QL réadapté, noté QL_r^m, qui remplace s par $s(t_i)$ dans l'équation (3.12) développée par Ibrahim et Whitt (2009c) pour les modèles M/M/s+GI. Ici $s(t_i)$ représente le nombre de serveurs dans le système au temps t_i où t_i désigne le temps estimé du prochain départ quand il reste i clients dans la file devant la nouvelle arrivée, et $t_{k+1} = t$. Les auteurs supposent qu'un événement de départ du système est soit une fin de service ou bien un abandon du client à la tête de la file. Chaque client dans la file a un temps de patience exponentiel de taux ψ_i dépendant de sa position i. Le taux total d'abandon quand il y a k clients dans la file est donné par $\alpha_k = \sum_{i=1}^{k} \psi_i$. Le temps entre le i-ième et $(i+1)$-ième départ du système est un exponentiel de taux $s\mu + \alpha_k - \alpha_{k-i}$. Ainsi le temps d'attente W d'un nouveau client ayant trouvé k autres clients dans la file est prédit par

$$\mathbb{E}[W_{\mathrm{QL}_r^m}(k)] = \sum_{i=1}^{k} \frac{1}{s(t_{i+1})\mu + \alpha_k - \alpha_{k-i}}, \tag{3.13}$$

et

$$\mathrm{t_i} = \mathrm{t_{i+1}} + \frac{1}{s\mu + \alpha_k - \alpha_{k-i}} \quad \text{pour} \quad 0 \leq i \leq k. \tag{3.14}$$

En pratique, dans les centres d'appels, il est souvent supposé que le nombre de serveurs et le taux d'arrivée sont variables entre les périodes, mais constant dans une période. Ibrahim et Whitt ont par la suite exploité les approximations des modèles fluides à trafic intense développées par Liu et Whitt (2010) pour obtenir un nouveau prédicteur pour le M(t)/M/$s(t)$+GI. Ce dernier est un prédicteur QL réadapté, appelé QL_r. Il prédit le temps d'attente du client qui entre dans la file

au temps t par

$$\mathbb{E}[W_{\mathrm{QL}_r}(k)] = v(t) \times \frac{k+1}{Q_f(t)+1} \tag{3.15}$$

où $v(t)$ est le temps d'attente potentiel du client et $Q_f(t)$ est la longueur de la file d'attente, qui sont estimés au temps t par les approximations fluides, et k est le nombre de clients réellement observé dans la file au temps t. Dans les grands systèmes, QL_r est plus précis que QL_r^m si la moyenne des durées de service est grande, mais devient moins précis si la moyenne des durées de service est petite.

3.2.2 Les prédicteurs pour les systèmes multi-compétences

Dans cette section, nous allons présenter les travaux qui sont effectués pour les systèmes multi-compétences. Une grande différence entre les centres d'appels multi-compétences et les centres d'appels à compétence unique (un seul type d'appel et un groupe d'agent) est l'importance jouée par la politique de routage dynamique dans les centres d'appels multi-compétences. Le routeur gère les files d'attente et affecte les appels aux agents disponibles. Avec une seule file d'attente FCFS, les futures arrivées n'affectent pas le temps d'attente des clients dans la file, mais ceci n'est pas nécessairement vrai dans le cas multi-compétences. Prédire les délais d'attente dans les centres d'appels multi-compétences est beaucoup plus difficile.

Il y a très peu de prédicteurs proposés pour les systèmes multi-compétences. De plus, tous les travaux existants avant les nôtres, à notre connaissance, sont faits pour des cas particuliers. Le seul qui peut s'utiliser dans n'importe quel système multi compétence est Ang et al. (2016) et il a été publié après les nôtres.

Les prédicteurs DH existants qui n'utilisent aucun paramètre du système peuvent être utilisés dans ce contexte. Cependant, les performances de ces derniers sont le plus souvent très mauvaises surtout s'il y a une variation dans les processus d'arrivée ou une variation du nombre de serveurs. Dans cette thèse, nous proposerons de nouveaux prédicteurs DH qui performent bien dans les systèmes multi-compétences. Ces nouveaux prédicteurs sont largement plus performants que les autres prédicteurs DH existants dans le cas multi compétences et sont très compétitifs par rap-

port aux prédicteurs QL dans les systèmes à file unique avec des agents identiques et des durées de service exponentielles. Cependant si les agents sont hétérogènes et les durées de service de loi log-normale, comme il est souvent le cas dans les données observées de la vie réelle, les prédicteurs que nous avons proposés sont largement plus performants que QL. La suite de cette section sera consacrée aux prédicteurs développés pour les systèmes multi-compétences.

Nakibly (2002) se concentre sur un système de file d'attente sans abandons avec deux types de clients et deux types de serveurs. Chaque agent a un ensemble de compétences et une liste de priorités dans le choix du type d'appel à traiter. Le prédicteur proposé utilise les paramètres et l'état du système pour estimer le délai d'attente d'un nouveau client. Plus précisément, l'auteur utilise une description de la chaîne de Markov du système avec les clients en attente pour estimer l'espérance du temps d'attente d'un nouveau client dans le système. Il montre aussi la complexité du calcul de l'estimation, et explique dans quelles conditions sa méthode est applicable. La méthode de Nakibly va être difficilement applicable dans les systèmes complexes de grande taille tels les centres d'appels actuels.

Armony et Maglaras (2004a, b) étudient un centre d'appels particulier avec deux types d'appels et un seul groupe d'agents qui est peut être représenté par un modèle V. Les appels de type 1 sont des appels réels et ceux de type 2 sont des appels virtuels qui doivent être servis avant d unités de temps (fixe). Ils considèrent qu'il n'y a pas d'abandons et le système est à trafic intense avec un grand nombre de serveurs. Les appels arrivent selon des processus de Poisson dont le taux λ_1 pour type 1 et λ_2 pour le type 2. L'espérance du temps W_1 d'attente d'un nouveau client ayant trouvé Q_1 autres clients en attente dans la file d'attente 1 est estimé par :

$$\mathbb{E}[W_1] = \frac{Q_1}{\lambda_1}. \tag{3.16}$$

Senderovich et al. (2015) ont développé des prédicteurs pour un système multi compétence particulier. Le modèle étudié a plusieurs types de clients et un seul groupe d'agents. Chaque agent du groupe peut servir tous les types d'appels. Pour

chaque type de client, les arrivées suivent un processus de Poisson dont le taux est variable dans le temps. Les durées de service et les temps de patience sont exponentiels de taux constant. Le nombre de serveurs est constant dans le temps. Les clients de même type sont toujours traités par la règle du *premier arrivé premier servi* et les clients de types différents sont traités selon une liste de priorité définie pour le groupe. Les auteurs ont proposé un prédicteur QL qui modifie légèrement le prédicteur développé par Whitt (équation 3.5). À chaque fin de service, le taux de sortie du système est ré-estimé en supposant que l'on connaît le type du prochain client à quitter le système. Ce prédicteur estime une borne supérieure et inférieure de l'espérance du temps d'attente pour chaque nouveau client. Ce prédicteur n'est pas utilisable dans les systèmes multi-compétences complexes, car dans ces systèmes, il est impossible de savoir les ordres de sortie des prochaines fins de service en observant l'état du système.

Un an après la publication de nos travaux sur la prédiction de délais pour les systèmes multi-compétences avec des algorithmes d'apprentissage machine (Thiongane et al., 2015), Ang et al. (2016) ont étudié la prédiction du temps d'attente dans les services d'urgence en utilisant avec un algorithme d'apprentissage différent de ceux que nous avons utilisés. Pour estimer le temps d'attente d'un patient, les auteurs utilisent un prédicteur qui combine la méthode du Lasso (Tibshirani, 1999) et la théorie des files d'attente. Ce prédicteur est appelé Q-Lasso pour faire référence à la méthode du Lasso et la théorie des files d'attente qu'ils utilisent. Q-Lasso prédit le temps d'attente d'un patient comme une fonction linéaire dépendant de l'état du système avec un objectif de minimiser le MSE des prédictions plus une fonction de pénalité pour éviter le sur apprentissage ("*over fitting*"). Leur définition de l'état du système à l'arrivée d'un patient est très similaire à la nôtre. Ici, en plus du LES, de la longueur de la file, du nombre de serveurs, de la période de la journée, les auteurs considèrent le prédicteur QL standard, le prédicteur QL des modèles fluides et la liste de priorité des serveurs dans la définition de l'état du système. Ang et al. utilisent les données de quatre services d'urgence pour montrer que Q-Lasso performe mieux que LES et ses variantes. Avec une légère réadaptation, ce

prédicteur pourrait bien être utilisé dans les centres d'appels multi-compétences. Nous allons comparer la précision de ce prédicteur avec les prédicteurs que nous avons proposé au chapitre 4 de cette thèse.

Résumé des travaux sur la prédiction et notre contribution.

D'après cette revue de la littérature, nous observons que plusieurs prédicteurs ont été développés. Cependant, nous notons qu'ils sont tous faits pour des systèmes particuliers. Ils ne performent pas bien dans tous les systèmes.

Les prédicteurs QL ne sont performants que si nous supposons que les agents sont identiques et les durées de service exponentielles. Dans les données de la vie réelle, nous observons que ces conditions ne sont pas souvent réunies. Les agents sont en général hétérogènes et les durées de service ne sont pas exponentielles comme on l'a traditionnellement supposé, mais plutôt de loi log-normale. À travers plusieurs exemples réalistes, nous observons que les prédicteurs QL fournissent de mauvais résultats.

Les prédicteurs DH, qui sont indépendants des paramètres des systèmes, sont utilisables dans n'importe quel type de système (file unique ou système multi-compétences). Mais on remarque que les performances de ces derniers sont très mauvaises quand il y a une variation dans le processus d'arrivée ou une variation du nombre de serveurs. Malheureusement, ces variations sont très présentes dans les systèmes de la vie réelle. Des réadaptations ont été proposées, mais les performances obtenues ne sont pas toujours satisfaisantes.

Dans cette thèse, nous avons proposé des réadaptations pour les prédicteurs DH qui donnent de très bons résultats dans tous les exemples que nous avons examinés. Avant cela nous avons aussi proposé des prédicteurs qui utilisent des idées de LES, de QL, et l'apprentissage machines. Ces prédicteurs sont indépendants du modèle et les performances enregistrées sont largement meilleures que celles des autres prédicteurs.

CHAPITRE 4

PRÉDICTEURS DE DÉLAIS POUR LES CENTRES D'APPELS MULTI-COMPÉTENCES BASÉS SUR L'APPRENTISSAGE MACHINE

4.1 Introduction

Dans ce chapitre, publié en partie dans Thiongane et al. (2015), nous développons des prédicteurs de délai pour les centres d'appels multi-compétences. Pour chaque type d'appel k, nous proposons un prédicteur qui prend en entrée le temps d'attente du dernier client de type j à entrer en service, la période d'arrivée du nouvel appel, le staffing des groupes d'agents, et les longueurs de file d'attente pour tous les types $i \neq k$ pour lesquelles il existe un agent qui peut servir les deux types k et i. Nous introduisons trois nouveaux prédicteurs qui utilisent la régression par les splines cubiques, les réseaux de neurones artificiels et le krigeage stochastique, respectivement, et dont les paramètres sont optimisés (ou appris) sur des données observées par simulation.

4.1.1 Objectifs

Notre étude se concentre sur la prédiction de délai dans les centres d'appels multi-compétences. Ceci est un problème important qui a à peine été étudié dans la littérature. Nos principaux objectifs sont les suivants : (i) tester la précision du prédicteur LES dans le cas multi-compétences, et (ii) proposer de nouveaux prédicteurs de délai qui peuvent rivaliser avec LES (éventuellement, être plus précis) dans le cas multi-compétence, et qui sont aussi précis ou plus précis que QL dans le cas des systèmes avec une seule file d'attente et un seul type d'appel. Dans le cas multi-compétence, nous ne considérons pas directement les prédicteurs QL, car ils ne s'étendent pas naturellement dans ce contexte. Ils auraient besoin de prendre en compte le partage de compétences des agents et la politique de routage, et

cela semble compliqué et difficile. Néanmoins, les nouveaux prédicteurs utilisent les longueurs des files d'attente en entrée, en combinaison avec d'autres informations.

Nos prédicteurs de délai proposés sont basés sur une approche heuristique qui combine l'approximation de fonctions, l'apprentissage machine, la simulation et des idées des prédicteurs de file d'attente unique. Pour chaque type d'appel k, le prédicteur est une fonction non linéaire paramétrée du délai attente du dernier client de type k qui est entré en service (comme dans LES), de la longueur de la file d'attente actuelle pour le type d'appel k, des longueurs de file d'attente pour tous les types $i \neq k$ pour lequel il existe un agent qui peut servir les deux types k et i, de la période d'arrivée de l'appel, et du staffing des groupes d'agents. Le vecteur de paramètre qui définit la fonction est "optimisé" (ou *appris*) pour minimiser l'erreur quadratique moyenne de prédiction, basé soit sur des données historiques réelles ou sur des données obtenues à partir d'une simulation du modèle de centre d'appel. Nos expériences portent sur ce dernier cas. Nous considérons trois types de fonctions de prédiction : (i) le premier type est défini par une régression par des splines de lissage, (ii) le second est défini par un réseau de neurone artificiel, et (iii) le dernier est défini par le krigeage stochastique. Quand un nouveau client entre dans la file d'attente k, la fonction de prédiction est évaluée, après avoir observé les entrées nécessaires. Les nouveaux prédicteurs peuvent être considérés comme des extensions de LES, ou des combinaisons partielles de LES et des prédicteurs basés sur QL. Ils nécessitent une étape supplémentaire "d'initialisation" (apprentissage).

Dans les expériences numériques, nos prédicteurs sont beaucoup plus précis que l'heuristique populaire LES (qui utilise, comme prédicteur, le délai du dernier client de même type qui a commencé son service) dans les cas multi-compétences. Dans les systèmes de file d'attente avec une seule file avec des agents homogènes, nos prédicteurs ont une précision très similaire à celle de QL (qui est le prédicteur optimal), mais dans le cas, le plus réaliste, des systèmes avec des serveurs hétérogènes, nos prédicteurs sont largement plus précis que QL.

4.1.2 Le plan du chapitre

Le reste du chapitre est organisé comme suit. La section 4.2 présente les nouveaux prédicteurs de délai proposés, les informations qu'ils utilisent en entrées, les paramètres des modèles, et explique comment ces paramètres sont estimés. Des expériences numériques pour différents modèles de centres d'appels sont présentées dans la section 4.3. Nous commençons en premier par trois centres d'appels modélisés par une seule file d'attente. Deuxièmement, nous utilisons deux modèles N de centres d'appels (avec deux types d'appels et deux groupes d'agents). Troisièmement, nous utilisons un modèle de centre multi-compétences basé sur des données réelles du centre d'appels d'un fournisseur de services publics au Québec. Le modèle comporte six catégories de clients (types d'appels), huit groupes d'agents, et les processus d'arrivée sont non-stationnaires. La section 4.4 étudie l'impact de l'ajout de nouvelles informations dans la définition de l'état du système. La section 4.5 présente les études de robustesse des prédicteurs. La section 4.6 compare nos prédicteurs à un autre prédicteur qui utilise l'apprentissage machine. Enfin, une conclusion et des remarques sont données dans la section 5.4.

4.2 Les prédicteurs de délai

4.2.1 Approximation de l'espérance conditionnelle du délai

Le temps d'attente $W > 0$ d'un client donné qui entre dans une file d'attente et attend jusqu'à ce que son service commence est une variable aléatoire dont la distribution dépend du type k de ce client et de l'état du système lorsque ce client arrive. Comme prédicteur simpliste de W, on peut tout simplement prendre le temps d'attente moyen global pour les clients de type k, qui peut être estimé par simulation (nous supposons qu'un modèle de simulation du système est disponible). Ce prédicteur est l'espérance inconditionnelle de W, $\mathbb{E}_k[W]$, lorsque nous prenons seulement k en entrée et nous ne regardons aucune autre information. Il est appelé le prédicteur NI ou *Non-Information predictor*. (Notez que nous avons défini W seulement pour un client qui entre dans la file d'attente et attend d'être servi, alors

l'attente est toujours conditionnelle à cela.)

Pour faire de meilleures prédictions, l'idée générale est d'observer l'état du système lorsque le client entre dans la file d'attente et retourner une estimation de l'espérance de W conditionnelle à cet état (étant donnée k). En pratique, nous allons sélectionner quelques informations $\boldsymbol{x}$ de l'état du système, et calculer une approximation de l'espérance conditionnelle $\mathbb{E}_k[W \mid \boldsymbol{x}]$, qui dépend de k. Cette approximation est définie par une *fonction de prédiction* $F_{k,\theta}(\boldsymbol{x})$ du vecteur d'information observé (entrée) $\boldsymbol{x}$, où θ est un vecteur de paramètres estimés (ou appris) précédemment.

Dans un premier temps, nous prenons $\boldsymbol{x} = (t, q, \boldsymbol{r})$ où t est le temps d'attente du dernier appel de type k à entrer en service, q est le nombre d'appels déjà dans la file k, et $\boldsymbol{r}$ est un vecteur qui contient la taille de chaque file d'attente $j \neq k$ tel qu'il y ait au moins un agent avec les deux compétences k et j.

Nous considérons trois façons de construire les fonctions $F_{k,\theta}$. Dans le premier cas, chaque fonction est une spline cubique lisse (régression des moindres carrés) qui est additive par rapport aux variables d'entrée (RS). Dans le second, la fonction est définie par un réseau de neurones artificiel multicouche (ANN). Dans le troisième cas, chaque fonction est définie par une régression de krigeage stochastique (SK). Ils sont décrits ci-dessous. Nous allons les comparer avec LES, avec le prédicteur simpliste NI (qui renvoie toujours le temps moyen d'attente comme une prédiction, ce qui correspond à prendre $\boldsymbol{x}$ vide) et aussi avec QL dans les cas où il est applicable.

Les prédicteurs sont optimisés pour minimiser l'*erreur quadratique moyenne* (MSE) des prédictions. Si $E = F_{k,\theta}(\boldsymbol{x})$ est le délai prédit pour un client "aléatoire" de type k qui opte pour attendre et W est son temps d'attente réalisé, nous rappelons que le MSE que nous approximons par contrepartie empirique, le ASE_k pour le type d'appels k est défini comme

$$\mathrm{MSE}_k = \mathbb{E}[(W - E)^2].$$

Pour estimer (ou apprendre) le vecteur de paramètres θ, nous utilisons un en-

semble de données d'apprentissage généré par simulation. Soit $\boldsymbol{x}_{k,c}$ représente le vecteur d'information $\boldsymbol{x}$ lorsque le c-ième appel (parmi C_k) de type k rejoint la file d'attente. Nous aimerions sélectionner θ pour minimiser le ASE_k tel que défini dans (2.2), avec $E_{k,c} = F_{k,\theta}(\boldsymbol{x}_{k,c})$. Cependant, d'autres facteurs peuvent également entrer dans la fonction objective ; par exemple les facteurs de lissage de la fonction de prédiction dans le cas des splines (voir ci-dessous).

4.2.2 Régression par des Splines de lissage (RS)

Les splines fournissent une classe bien connue et puissante de méthodes d'approximation pour les fonctions générales lisses de Boor (1978). Ici, nous utilisons les splines lisses cubiques, dont les paramètres sont estimés par régression des moindres carrés avec un terme de pénalité sur la variation de la fonction, afin de promouvoir les fonctions lisses. Nous nous limitons également aux splines additives, qui peuvent être écrites comme une somme de fonctions unidimensionnelles. Autrement dit, si le vecteur d'information est écrit comme $\boldsymbol{x} = (x_1, \ldots, x_D)$, le prédicteur spline additif peut être écrit comme

$$F_{k,\theta}(\boldsymbol{x}) = \sum_{d=1}^{D} f_d(x_d),$$

où chaque f_d est une spline cubique à une dimension. Les paramètres de toutes ces fonctions splines f_d forment le vecteur θ. Ces paramètres doivent satisfaire les contraintes que les morceaux successifs de la spline (qui sont des polynômes cubiques) ont leurs dérivées premières et secondes égales aux bornes. Pour estimer les paramètres, nous utilisons la fonction `gam` mis en oeuvre dans le package `mgcv` du logiciel statistique R (R Core Team, 2014, Wood, 2006). Le nombre de points de noeuds et les facteurs de lissage sont choisis automatiquement par le package, en fonction des données.

4.2.3 Les réseaux de neurones artificiels (ANN)

Les réseaux de neurones artificiels ou "Artificial Neural Networks" (ANNs) sont un autre moyen très populaire et efficace pour approximer des fonctions complexes

de grande dimension. Une tendance récente est l'*apprentissage profond*, qui se réfère à l'utilisation des ANNs avec plusieurs couches de neurones Bengio et al. (2012), et LeCun et al. (2015). Nous adoptons cette technologie ici. Pour entraîner le réseau de neurones (c'est à dire estimer un bon vecteur de paramètres θ), nous utilisons le logiciel Pylearn2 (Goodfellow et al., 2013). Nous avons sélectionné un réseau de neurones multicouche dans lequel les sorties des noeuds à la couche l sont les entrées de chaque noeud à la couche suivante $l+1$. Un ANN typique a une couche d'entrée, une couche de sortie et plusieurs couches cachées. Il n'existe actuellement aucune méthode pour déterminer le nombre optimal de couches cachées ou nombre de noeuds en eux. En pratique, ces valeurs sont généralement choisies, par essais et erreurs, après quelques essais préliminaires. Dans nos exemples numériques, nous utilisons cinq couches ou plus. Dans le cas où le nombre de couches est cinq, nous avons une couche d'entrée, trois couches cachées et une couche de sortie. Le nombre de noeuds dans la couche d'entrée est égal au nombre d'éléments dans le vecteur de paramètres $\boldsymbol{x}$, et la couche de sortie a un seul noeud qui renvoie le délai estimé. Le nombre de noeuds dans une couche cachée dépend de la taille du centre d'appels ; ce nombre est spécifié dans la section numérique. Pour chaque noeud caché, nous utilisons une fonction de transfert appelée "*rectifier activation function*", $h(\boldsymbol{z}) = \max(0, b+\boldsymbol{w}\cdot\boldsymbol{z})$, où $\boldsymbol{z}$ est le vecteur des entrées pour le noeud, tandis que la constante b et le vecteur des coefficients $\boldsymbol{w}$ sont des paramètres appris durant l'entraînement. Le (grand) vecteur θ contient l'ensemble de tous ces paramètres b et $\boldsymbol{w}$, sur tous les noeuds. Ce type de fonction d'activation a été proposé récemment par Glorot et al. (2011), et on pense actuellement qu'il représente plus fidèlement le mécanisme biologique d'un neurone que les fonctions classiques sigmoïde et tangente hyperboliques. Les paramètres sont appris par un algorithme de rétropropagation qui utilise une méthode de descente du gradient (Bishop, 2006).

Ces ANNs sont très puissants, mais un inconvénient est qu'ils requièrent de grands échantillons d'entraînement et leur entraînement peut prendre beaucoup plus de temps que pour les autres techniques de régression telles que les splines. Pour accélérer l'apprentissage, nous utilisons l'agrégation de données, comme suit. L'idée

est de regrouper les observations, dont les valeurs, de $\boldsymbol{x}$ sont presque les mêmes, et les remplacer par une seule observation $(\boldsymbol{x}', w')$, où w' est le temps d'attente moyen pour les observations qui ont été agrégées. Cette nouvelle observation agrégée aura un poids proportionnel au nombre d'observations originales qui ont été regroupées en elle. Pour former les groupes qui sont agrégés, nous regroupons en premier toutes les observations $\boldsymbol{x}$ ayant la même paire $(q, \boldsymbol{r})$, puis divisons chacun de ces groupes en 20 sous-groupes de taille à peu près égale en fonction de la valeur de t. Pour cela, nous utilisons le 5%, 10%, 15%, ..., quantiles par rapport à t comme séparateurs pour faire les sous-groupes, puis agréger chaque sous-groupe. L'observation agrégée est $(\boldsymbol{x}', w') = ((t', q, \boldsymbol{r}), w')$, où w' est le temps d'attente moyen pour le sous-groupe et t' est le milieu de l'intervalle entre les quantiles correspondants. L'ensemble des observations agrégées est utilisé comme le nouvel ensemble de données d'entraînement.

4.2.4 Le krigeage stochastique (SK)

Les méthodes de régression, comme RS, supposent que les données observées (les délais) W, peuvent être modélisées par une fonction déterministe f plus un certain bruit stochastique ϵ :

$$W(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon(\boldsymbol{x}).$$

Ce bruit stochastique est appelé bruit intrinsèque, car il est inhérent au modèle stochastique. La méthode du krigeage modélise la surface de réponse W pour les données non observées (les délais non observés) par :

$$W(\boldsymbol{x}) = f(\boldsymbol{x}) + M(\boldsymbol{x}),$$

où $M(\boldsymbol{x})$ est un bruit extrinsèque avec une corrélation spatiale basée sur des données observées. L'incertitude extrinsèque n'est pas une propriété du modèle lui-même, mais une description de notre incertitude du temps d'attente à un point x où nous

n'avons pas observé de données. Le Kriging suppose qu'il n'y a pas de bruit sur les données observées. $M(\boldsymbol{x})$ et $M(\boldsymbol{x}')$ auront tendance à être similaire si $\boldsymbol{x}$ et $\boldsymbol{x}'$ sont proches l'une de l'autre dans l'espace, et leur corrélation dépend de la différence $\boldsymbol{x} - \boldsymbol{x}'$. La principale différence entre la "Régression" et le "Krigeage" est la suivante : le Krigeage traite la surface de réponse comme un champ Gaussien alors que la régression traite la surface de réponse comme une fonction déterministe et l'erreur comme une variable aléatoire. Le krigeage stochastique (SK) Ankenman et al. (2010), Staum (2009) utilise les bruits intrinsèque $\epsilon(\boldsymbol{x})$ et extrinsèque $M(\boldsymbol{x})$ pour améliorer le modèle de prédiction :

$$F_{k,\boldsymbol{\theta}}(\boldsymbol{x}) = f(\boldsymbol{x}) + M(\boldsymbol{x}) + \epsilon(\boldsymbol{x}). \tag{4.1}$$

La fonction de prédiction a une forme linéaire qui dépend des matrices de covariance de $M(\boldsymbol{x})$ et $\epsilon(\boldsymbol{x})$. Le SK fournit une meilleure prédiction lorsque les deux sources d'incertitude sont non-négligeables. Les paramètres de toutes ces fonctions $f(\boldsymbol{x})$, $M(\boldsymbol{x})$, $\epsilon(\boldsymbol{x})$, forment le vecteur θ.

Pour entraîner le prédicteur dans le cas du SK, pour chaque $\boldsymbol{x}$ nous avons besoin de la moyenne et de la variance du temps d'attente W mais aussi des matrices de covariance intrinsèque et extrinsèque. Nous agrégeons les données de la même façon que le cas des réseaux de neurones (section 4.2.3) pour obtenir l'ensemble $(\boldsymbol{x}', w') = ((t', q, \boldsymbol{r}), w')$. Nous ajoutons à cet ensemble la variance empirique de w' que nous notons par v' pour obtenir l'ensemble d'entraînement $(\boldsymbol{x}', w', v') = ((t', q, \boldsymbol{r}), w', v')$. Les matrices intrinsèque et extrinsèque sont déterminées par la fonction `mlegp`, mise en oeuvre dans le package `mlegp` du logiciel statistique R (G. M. Dancik, 2015, R Core Team, 2014), avant le début de l'optimisation de θ.

4.3 Expériences numériques

Nous comparons les performances et la précision des prédicteurs LES, QL, RS, SK et ANN en simulant de petits et grands modèles de centres d'appels. Dans la première partie de cette étude numérique, nous utilisons des systèmes avec seule

file d'attente et un seul type d'appel. Nous commençons avec les files classiques M/M/s et M/M/s+M pour lesquels nous avons des formules analytiques pour l'espérance du délai d'attente dépendant de l'état du système. Nous supposons que les s agents sont identiques et ont la même distribution exponentielle pour les durées de service. Dans ces conditions, les prédicteurs QL sont optimaux. Notre but est de vérifier si nos prédicteurs RS, SK et ANN sont compétitifs avec les meilleurs prédicteurs disponibles pour ces modèles. Ensuite, nous considérons des modèles plus réalistes dans le contexte des centres d'appels : le modèle M/M/s+M avec des agents hétérogènes avec des durées exponentielles ; le modèle M/LN/s+M avec des agents hétérogènes et des durées de service de loi log-normale. Chaque agent a une distribution pour les temps de service avec ses propres paramètres. Dans ces nouvelles conditions, QL est utilisable, mais il n'est pas le prédicteur optimal. L'objectif est de voir si nos prédicteurs pourront faire mieux que QL.

Dans la seconde partie, nous testons nos prédicteurs sur des modèles de centres d'appels multi-compétences. Nous commençons par deux modèles N de centres d'appels, l'un avec des files d'attente courtes et l'autre avec de longues files d'attente. Nous terminons par un modèle de centre d'appels multi-compétences plus grand et plus complexe basé sur des données réelles (HQ). Le modèle comporte six types d'appels et huit groupes d'agents.

Comme expliqué dans la section 4.2.1, les prédicteurs sont optimisés pour minimiser le ASE, mais nous prenons la valeur de la racine normalisée, le RRASE. Pour entraîner les prédicteurs RS, SK et ANN, nous générons les observations sur les appels par simulation. Pour évaluer le M/M/s et M/M/s+M à l'état d'équilibre, nous simulons une longue exécution de 600,000 heures, et les observations des premières 200,000 heures (temps de préchauffage) sont éliminées. Pour les modèles N et HQ, nous simulons 100 jours indépendants pour générer les observations. Pour l'ANN, nous prenons 80% des observations comme données d'entraînement et les 20 % restants comme données de "*test*" et "*validation*" utilisées pour sélectionner le meilleur paramètre θ parmi ceux trouvés durant l'entraînement. Pour comparer le RRASE des différents prédicteurs, nous générons un ensemble indépendant

d'observations en exécutant une autre simulation de même longueur, pour chaque modèle, et nous utilisons ce même ensemble pour calculer le RRASE pour tous les prédicteurs, de sorte que les prédicteurs sont effectivement comparés sur les mêmes données.

4.3.1 Modèles à file unique avec des agents homogènes et durées de service exponentielles

Nous comparons nos prédicteurs avec LES et QL dans des centres d'appels qui sont modélisés par des systèmes M/M/s et M/M/s+M. Nous considérons que les agents sont homogènes et ont la même distribution des temps de service. Comme nous l'avons dit un peu plus tôt, les prédicteurs QL sont optimaux dans ces systèmes à l'état d'équilibre. Nous avons étudié chacun des systèmes dans différents types de régimes.

4.3.1.1 Expériences avec un modèle M/M/s

Nous considérons une seule file d'attente avec un processus d'arrivée de Poisson de taux λ appels par heure, les temps de service exponentiels de moyenne $\mu^{-1} = 2$, et on a s agents identiques. Il n'y a pas d'abandons. Dans les simulations utilisées pour évaluer les prédicteurs, nous avons essayé plusieurs exemples de différente taille et charge (du système) $\rho = \lambda/s\mu$. Le premier exemple à une file d'attente courte : la moitié des clients est servie immédiatement à leur arrivée et l'autre moitié a eu à attendre avant d'être servi. Le nombre d'agents est $s = 26$ et la charge du système $\rho = 0.9$. Le second exemple a le même nombre d'agents que le premier, mais une charge un peu plus grande, $\rho = 0.96$. Cet exemple a un taux d'arrivée plus grand et une file d'attente plus longue que le premier. Notre troisième exemple est un grand système dans un régime à trafic intense ($s = 100$ et $\rho = 0.99$). Le tableau 4.1 donne les mesures de performances des trois centres d'appels.

Le tableau 4.2 montre le RRASE obtenu pour chaque prédicteur dans les trois exemples. Comme prévu, QL a les meilleurs résultats, car c'est un prédicteur de

Paramètres		Performances		
ρ	s	PD	AWT (sec.)	AQS
0.90	26	51.4	372	4.8
0.96	26	78.2	1795	19.5
0.99	100	87.9	1596	87.7

Tableau 4.1 : Les performances des différentes modèles M/M/s utilisés.

Paramètres		Prédicteurs					
ρ	s	NI	LES	QL	RS	SK	ANN
0.90	26	0.994	0.449	**0.308**	0.310	0.315	0.310
0.96	26	0.956	0.266	**0.244**	0.256	0.245	0.258
0.99	100	0.996	0.141	**0.099**	0.101	0.108	0.102

Tableau 4.2 : Les RRASEs pour le modèle M/M/s.

délai optimal pour une file d'attente M/M/s à l'état d'équilibre. RS, SK et ANN arrivent justes derrière QL, suivis par LES. Les précisions de QL, RS, SK et ANN sont tout à fait comparables, avec une différence inférieure à 1%. LES a une RRASE plus élevée, d'environ 12%. NI donne le pire résultat, sans surprise, ce qui confirme que l'utilisation de prédicteurs dépendant de l'état vaut la peine. Nous avons testé une variante de RS et de SK où l'entrée est seulement q (le nombre de clients déjà en file d'attente) et une variante de ANN à laquelle nous donnons la même entrée que le prédicteur QL, qui est la longueur de la file d'attente q, et les constantes μ et s. Ces variantes ont également donné des résultats sensiblement égaux à ceux de QL et montrent que la fonction de prédiction QL peut être apprise avec des méthodes d'apprentissage machine.

4.3.1.2 Modèle M/M/s+M

Nous considérons maintenant un centre d'appels avec une seule file d'attente de capacité infinie avec abandons. Les appels arrivent au centre selon un processus de Poisson de taux λ appels par heure. Les durées de patience suivent une distribution exponentielle de moyenne $\nu^{-1} = 0.6$ heure. Les durées de service suivent une distribution exponentielle de moyenne $\mu^{-1} = 0.5$. Le nombre d'agents qui répondent

aux appels est s et nous supposons qu'ils sont tous identiques. Comme dans le cas de la file sans abandon, nous avons encore utilisé trois exemples de centres. Le premier avec courte file d'attente a un nombre d'agents $s = 17$, et un $\rho = 1.11$. Dans ce système, plus de 70% des appels servis ont un temps d'attente strictement positif. Le second a toujours le même nombre de serveurs que le précédent et a une charge ρ un peu plus élevé de 1.35. Nous observons une longueur de file d'attente moyenne deux fois plus grande que celui du premier exemple. Le troisième est un grand système dans un régime à trafic intense avec un $\rho = 1.5$ et un $s = 100$. Le tableau 4.3 affiche les mesures de performances des différents centres.

Paramètres		Performances			
ρ	s	PD	PA	AWT (sec.)	AQS
1.11	17	70.9	15.23	274.7	2.9
1.35	17	91.6	26.9	484.5	6.2
1.50	100	99.9	33.3	596	49.9

Tableau 4.3 : Les performances des différents modèles M/M/s+M utilisés.

Le tableau 4.4 montre le RRASE obtenu pour chaque prédicteur. Encore pas de surprise, le prédicteur optimal QL donne les meilleurs résultats. RS, SK et ANN suivent de très près QL. Ces derniers sont suivis par LES. NI arrive en dernière position. Comme pour le cas précédent, les précisions de QL, RS, SK et ANN sont tout à fait comparables, avec une différence inférieure à 1%. Les nouveaux prédicteurs RS, ANN et SK ont des RRASEs beaucoup plus petits que celui de LES. La différence de précision diminue quand ρ augmente. Nous constatons encore que NI est le prédicteur le plus mauvais. Dans tous ces exemples, nous observons que les prédicteurs qui utilisent l'apprentissage machine ont bien appris la fonction prédiction et ont des performances très similaires à celles de QL. Dans nos prochains exemples, nous n'allons plus utiliser NI.

Paramètres		Prédicteurs					
ρ	s	NI	LES	QL	RS	SK	ANN
1.11	17	0.998	0.704	**0.466**	0.468	0.460	**0.466**
1.35	17	0.989	0.568	**0.380**	0.385	0.384	0.383
1.50	100	0.984	0.204	**0.143**	0.144	0.144	**0.143**

Tableau 4.4 : Les RRASEs pour le modèle M/M/s+M.

4.3.2 Modèles à file unique avec des agents hétérogènes et des durées de service exponentielles

Nous considérons une file d'attente M/M/s+M avec des agents hétérogènes. Nous supposons qu'il y a un seul type de client et les agents qui répondent aux appels sont hétérogènes. Bien que cette hétérogénéité est souvent ignorée, des investigations empiriques des données recueillies dans les centres d'appels de la vie réelle montrent que les agents possèdent des distributions du temps de service différentes pour un même type d'appel (Gans et al., 2010, Ibrahim et al., 2016b). Notre objectif est de comparer la précision et les performances de nos prédicteurs à celles de QL et de LES dans ce cas de figure.

Dans les exemples qui seront étudiés dans cette section, les arrivées suivent un processus de Poisson de taux λ, les temps de patience sont exponentiels de moyenne ν^{-1}, et on a s agents qui répondent aux appels. Les temps de service sont exponentiels et chaque agent i a une distribution du temps de service distincte de moyenne μ_i^{-1}. Dans les expériences numériques, nous prenons les données d'un vrai centre d'appels présentées dans Gans et al. (2010). Dans ces données, nous avons les moyennes des temps de service de 12 agents hétérogènes pour un type d'appel A. On suppose que les temps de service sont exponentiels. Les temps de patience pour ce type d'appel A sont exponentiels de moyenne $\nu^{-1} = 0.5$ heures. Le tableau 4.5 rapporte les moyennes μ^{-1} observées dans ces données pour ces 12 agents. Dans nos simulations, nous utilisons souvent un staffing supérieur à 12 agents et dans ces cas nous supposons que des agents clones existent pour nous permettre d'avoir le staffing voulu.

Nous avons étudié trois exemples de centres d'appels. Le premier centre (C1)

Agents	1	2	3	4	5	6	7	8	9	10	11	12
$1/\mu$	0.259	0.247	0.218	0.216	0.215	0.208	0.207	0.199	0.186	0.173	0.159	0.158

Tableau 4.5 : Le taux de service et moyenne du temps de service des 12 agents pour le type d'appel A.

est de petite taille. Il a 12 agents et un taux d'arrivée $\lambda = 100$. Le taux de service des agents sont ceux des 12 agents du tableau 4.5. Le second centre (C2) est de taille moyenne avec 36 agents a pour taux d'arrivée $\lambda = 300$. Nous supposons qu'il existe trois clones de chacun des agents pour avoir 36 agents. Le troisième centre (C3) est de grande taille avec 240 agents. Nous supposons qu'il est dans un régime à trafic intense. Son taux d'arrivée $\lambda = 2000$ et nous supposons qu'il existe 20 agents clones de chacun des 12 agents pour satisfaire le staffing du modèle. Le tableau 4.6 donne les mesures de performances de ces trois centres d'appels.

	Mesures de performances			
Centres d'appels	PD (%)	PA(%)	AWT(sec.)	AQS
C1	97.9	37.7	666.8	18.6
C2	98.1	37.6	676.0	55.7
C3	98.2	37.5	676.4	371.3

Tableau 4.6 : Les mesures de performances pour les modèles M/M/s+M

Dans la suite, nous allons comparer les performances des prédicteurs QL, LES, RS, SK et ANN dans les 3 centres d'appels. Mais avant cela, nous décrirons, une manière d'utiliser le prédicteur QL dans un système avec des agents hétérogènes.

Utilisation du prédicteur QL dans un modèle M/M/s+M avec des serveurs hétérogènes

Les prédicteurs QL sont développés pour des systèmes de file d'attente avec des serveurs identiques. Ils utilisent la longueur k de la file, le nombre s de serveurs dans le système, le taux d'abandon ν et le taux moyen de service des agents μ pour prédire le temps d'attente des clients. Pour ces prédicteurs, on suppose qu'il existe une seule distribution exponentielle des temps de service de taux μ pour l'ensemble des agents. Dans un système avec agents hétérogènes, nous pouvons faire une abs-

traction de cette hétérogénéité des agents et utiliser QL avec le taux de service moyen de l'ensemble des agents que nous appelons $\bar{\mu}$. Bien sûr, nous supposons que tous les agents sont occupés tant qu'il y a des clients en attente dans la file. Une fois la valeur de $\bar{\mu}$ estimée, nous prédisons l'espérance du temps d'attente W d'un client ayant trouvé k autres clients dans la file par

$$\mathbb{E}[W] = \sum_{i=1}^{k} \frac{1}{s\bar{\mu} + i\nu}.$$

Pour estimer $\bar{\mu}$ nous pouvons utiliser deux méthodes. Parfois, dans les données disponibles aux centres d'appels, l'information qui est enregistrée sur les durées de service d'un agent c'est la moyenne journalière de ces temps de service. Voici la description des deux méthodes en utilisant ces informations :

Méthode 1 : Nous calculons la moyenne des moyennes des temps de service de l'ensemble des agents et ensuite nous inversons cette moyenne pour trouver $\bar{\mu}$.

Méthode 2 : Nous calculons le taux de service de chaque agent en inversant sa moyenne des temps de service puis nous calculons le taux de service moyen du système par la moyenne des taux de service des agents.

Plusieurs exemples étudiés suggèrent que la Méthode 2 est toujours meilleure que la Méthode 1. Voici un petit exemple qui illustre ceci.

Prenons un petit centre d'appels avec un seul type d'appel et deux agents, a_1 et a_2. Supposons que le centre est ouvert pendant 10 heures de temps dans la journée et que la file d'attente du centre n'est jamais vide (de l'ouverture à la fermeture, il y a toujours des clients en attente de service). Supposons aussi que l'agent a_1 est rapide pour servir les appels avec un taux de service $\mu_1 = 50$ appels par heure, et que l'agent a_2 est lent avec un taux de service $\mu_2 = 25$ appels par heure. Le taux service moyen obtenu avec la Méthode 1 (moyenne sur les temps) est $\bar{\mu} = 33.3$ et celui donné par la Méthode 2 (moyenne sur les taux) est $\bar{\mu} = 37.5$.

Dans la journée de 10 heures, l'agent a_1 va servir en tout $10 \times 50 = 500$ appels, et l'agent a_2 quant à lui servira $10 \times 25 = 250$. Le taux de service moyen réel par heure dans la journée est $[(500 + 250)/2]/10 = 37.5$. Il correspond bien au taux de service moyen donné par la "Méthode 2". Cet exemple est l'un parmi plusieurs que nous avons essayés qui montre que la Méthode 2 est meilleure que la Méthode 1 pour déterminer le taux de service moyen dans un système avec des agents hétérogènes.

Nous avons aussi évalué le RRASE du prédicteur QL dans les trois exemples de centres d'appels (C1, C2 et C3) avec des $\bar{\mu}$ déterminés par les Méthode 1 et 2. Ces prédicteurs QL sont respectivement notés QL(Méthode 1) et QL(Méthode 2). Le tableau 4.7 présente les RRASEs de QL(Méthode 1) et QL(Méthode 2). Dans tous les exemples, nous observons QL(Méthode 2) donne les meilleurs résultats. Ces résultats suggèrent qu'il est préférable de faire la moyenne sur les taux que de faire la moyenne sur les temps. Dans la suite de ce document, nous utiliserons toujours QL comme étant QL(Méthode 2) lors de l'étude d'un système avec des agents hétérogènes.

	RRASE	
Centres	QL(Méthode 1)	QL(Méthode 2)
C1	0.249 ±0.001	0.239 ±0.001
C2	0.160 ±0.001	0.153 ±0.001
C3	0.101 ±0.001	0.075 ±0.001

Tableau 4.7 : Les RRASEs des centres d'appels C1, C2 et C3 avec les prédicteurs QL(Méthode 1) et QL(Méthode 2).

Nous simulons 100 journées indépendantes des centres d'appels C1, C2 et C3 et dans chacun des centres d'appels, nous évaluons le RRASE des prédicteurs LES, QL, RS, ANN et SK . Le tableau 4.8 rapporte les RRASEs des prédicteurs dans les trois exemples. Contrairement, à l'exemple avec des agents homogènes, nous constatons dans ces trois exemples que les nouveaux prédicteurs (ANN, RS et SK qui ont des résultats très similaires) sont un peu plus précis que QL. La différence de précision entre les nouveaux prédicteurs et QL varie entre 1 et 3%. LES donne toujours les plus mauvais résultats.

La perte de précision de QL s'explique par l'hétérogénéité des agents qui fait que les durées de service des appels proviennent de diverses distributions exponentielles et non d'une seule et unique distribution exponentielle. Les prédicteurs RS, ANN et SK quant à eux ont appris les bons paramètres de la fonction à travers les données historiques.

Bien que les résultats de QL sont moins précis que ceux des nouveaux prédicteurs, il reste toujours un bon prédicteur dans ces systèmes. Cependant, si des données détaillées sur les agents ne sont pas disponibles (dans ce cas, on ne peut plus estimer le meilleur $\bar{\mu}$ pour QL qui dépend des agents sélectionnés), alors QL peut donner de mauvais résultats.

	RRASE				
Centres	LES	QL	RS	SK	ANN
C1	0.344	0.239	0.221	0.223	**0.222**
C2	0.204	0.153	0.127	0.132	**0.124**
C3	0.087	0.075	0.052	0.055	**0.050**

Tableau 4.8 : Les RRASEs pour les modèles M/M/s+M.

4.3.3 Modèles à file unique avec des agents hétérogènes et des durées de service de loi log-normale

Nous considérons une file d'attente M/LN/s+M avec des serveurs hétérogènes et des temps de service de loi log-normale basé sur des données réelles. Des études empiriques effectuées sur des données réelles ont montré qu'en plus de l'hétérogénéité des agents, les durées de service sont de loi log-normale plutôt qu'exponentielles (Brown et al., 2005, Ibrahim et al., 2016b, Shen et Brown, 2006).

Nous étudions un exemple dans lequel les agents utilisés sont ceux d'un véritable centre d'appels. Ces agents présentés dans le supplément en ligne de Ibrahim et al. (2016b) sont au nombre de 10 et ils répondent a appel de type F. Le tableau 4.9 rapporte la moyenne et la variance des durées de service des agents. Elles sont estimées sur des données collectées sur 45 semaines. Chaque agent i a distribution des temps de service de loi log-normale de moyenne m_i et de variance v_i. Les

paramètres d'échelle κ_i et de forme σ_i sont déterminés par les formules de Mood et al. (1974) pages 540–541 :

$$\kappa_i = \ln\left(\frac{m_i}{\sqrt{1+\frac{v_i}{{m_i}^2}}}\right) \quad et \quad \sigma_i = \sqrt{\ln\left(1+\frac{v_i}{{m_i}^2}\right)}. \tag{4.2}$$

La journée est divisée en 10 périodes d'une heure. Dans chaque période p, le

Agents	1	2	3	4	5	6	7	8	9	10
m_i (heure)	0.186	0.080	0.099	0.0345	0.0452	0.196	0.1117	0.133	0.126	0.0474
v_i	0.055	0.018	0.0248	0.0009	0.0027	0.157	0.01667	0.022	0.170	0.0034

Tableau 4.9 : La moyenne m et la variance v des temps de service des agents pour le type d'appel B.

processus d'arrivée est Poisson avec un taux constant. Nous prenons $\lambda_p = 320$ pour p impair et $\lambda_p = 200$ pour p pair. Les temps de patiences sont des exponentiels de moyenne $\nu^{-1} = 2$.

Pour évaluer le RRASE des prédicteurs LES, QL, RS, SK et ANN, nous simulons 100 journées indépendantes du centre d'appels. Nous observons que la longueur moyenne de la file d'attente est de 37.4, la proportion de délais est de 98.2%, la proportion d'abandons de 29.5%, et le temps d'attente moyen des clients dans la file est de 529.2 secondes.

Le tableau 4.10 rapporte les RRASE×100 des différents prédicteurs. Comme dans l'exemple précédent (avec agents hétérogènes et temps de service exponentiels), nous observons que RS, SK et ANN donnent les meilleurs résultats. Mais cette fois-ci, la différence de précision entre ces derniers et QL est très élevée (environ 15%). LES et QL donnent des résultats très similaires (LES est plus d'environ 1%). Dans cet exemple et comme dans plusieurs autres exemples que nous avons étudiés, nous observons que les prédicteurs QL sont largement moins précis que les prédicteurs RS, SK et ANN pour les systèmes de file d'attente avec des agents non identiques et des durées de service de loi log-normale.

Dans la suite de cette section, nous allons étudier trois exemples de centres d'appels multi-compétences. Les prédicteurs QL ne sont pas applicables dans ces

	LES	RS	SK	ANN	QL
RRASE	30.6	16.5	16.1	**14.3**	31.8

Tableau 4.10 : RRASE des prédicteurs pour le modèle M/LN/10+M avec des taux d'arrivée variable dans le temps.

systèmes. Dans les deux premiers exemples, nous considérons des modèles N avec agents homogènes et des durées de service exponentielles. Pour le troisième exemple basé sur des données réelles, les agents sont hétérogènes avec des durées de service de loi log-normale.

4.3.4 Modèles N

Nous considérons maintenant un modèle N. Une journée est divisée en $P = 10$ périodes de 1 heure. Les processus d'arrivée sont des processus de Poisson de taux constant dans chaque période. Toutes les durées de service et les temps de patience sont exponentiels et indépendants. Nous utilisons une politique de *routage par priorité* (Chan et al., 2014) qui fonctionne comme suit. Pour un appel de type 1, le routeur va d'abord essayer de l'affecter à un agent libre du groupe 1. S'il n'y a pas d'agent libre, alors le routeur va essayer de l'attribuer à un agent libre du groupe 2. Les agents du groupe 2 donnent toujours la priorité aux appels de type 2, même si certains appels de type 1 ont attendu plus longtemps. Ainsi, les appels du même type sont de premier arrivé, premier servi, mais les appels de différents types peuvent être servis dans des ordres différents.

Nous testons les prédicteurs sur deux instances du modèle très différentes : (i) des files d'attente et des temps d'attente courts, et (ii) des longues files d'attente et temps d'attente. Ces deux systèmes se comportent très différemment parce que très peu d'agents du groupe 2 servent des appels de type 1 lorsque les files d'attente sont (presque) toujours remplies. Asymptotiquement, aucun agent du groupe 2 ne devrait servir des appels de type 1 parce que la file d'attente 2 n'est jamais vide. Avec les files d'attente courtes, le système a beaucoup plus de variabilité et son étude est plus intéressante. Si les files d'attente n'étaient jamais vides, aucun des

agents du groupe 2 ne servirait des appels de type 1, et nous aurions deux files d'attente simples séparées.

Nous comparons les prédicteurs LES, RS, SK et ANN. Nous n'incluons pas QL parce qu'il n'est pas directement applicable (il n'y a pas de formule) dans le contexte multi-compétences. L'entrée pour RS, SK et ANN est $\boldsymbol{x} = (t, q, \boldsymbol{r})$ pour les deux types d'appels, où $\boldsymbol{r}$ est un vecteur de longueur 1. Pour l'ANN, nous avons constaté que 180 noeuds par couche cachée étaient suffisants. Nous considérons également des variantes de RS, SK et ANN où la taille de la file d'attente secondaire (l'entrée $\boldsymbol{r}$) est retirée de l'entrée de prédiction. Nous nommons ces variantes RS(t, q), SK(t, q) et ANN(t, q). Dans chaque cas, nous rapportons le RRASE pour les appels de type 1, de type 2, et l'agrégation des deux types.

4.3.4.1 Un modèle N avec courtes files d'attente

Notre premier exemple de modèle N est un cas avec des files d'attente courtes. Pour les appels de type 1, le vecteur des taux d'arrivées (par période) est $\lambda_1 = (16, 20, 28, 30, 35, 45, 40, 30, 20, 15)$ par heure, la moyenne du temps de service $\mu_1^{-1} =$ 20 minutes, et la moyenne du temps de patience est $\nu_1^{-1} = 25$ minutes. Pour le type 2, les taux d'arrivées sont $\lambda_2 = (20, 32, 40, 50, 60, 50, 40, 35, 30, 20)$ par heure, le temps de service moyen est de $\mu_2^{-1} = 10$ minutes, et la moyenne des temps de patience est $\nu_2^{-1} = 20$ minutes. Les vecteurs de staffing (par période) sont $s_1 = (3, 5, 8, 8, 9, 10, 9, 6, 5, 5)$ et $s_2 = (4, 6, 8, 10, 9, 9, 8, 8, 6, 5)$.

Les mesures de performance agrégées sur toutes les périodes sont les suivantes. Pour l'appel de type 1, il y a une moyenne de 1.6 client dans la file d'attente, la proportion de délais est de 61%, la proportion d'abandons est de 14%, le temps d'attente moyen est de 211 secondes (pour tous les clients), et le temps d'attente moyen des clients qui ont attendu et ont été servis était de 354 secondes. Pour le type 2, ces mesures moyennes sont respectivement de 2.9 clients dans la file d'attente, 79% attendus, 12% d'abandons, 193 secondes d'attente, et 248 secondes d'attente. Le partage des compétences est très présent : 80% des appels de type 1 servis ont été répondu par des agents du groupe 1, tandis que l'autre 20% ont été

répondu par des agents du groupe 2. Bien que les temps d'attente moyens globaux (sur tous les clients) diffèrent de moins de 20 secondes entre les deux types d'appels, les temps d'attente moyens pour ceux qui ont attendu et ont été servis sont très différents pour les deux types. Figure 4.1 montre la distribution de temps d'attente des clients qui ont attendu et ont été servis, pour chaque type. Le type 1 a une file plus longue et évidemment une grande variance.

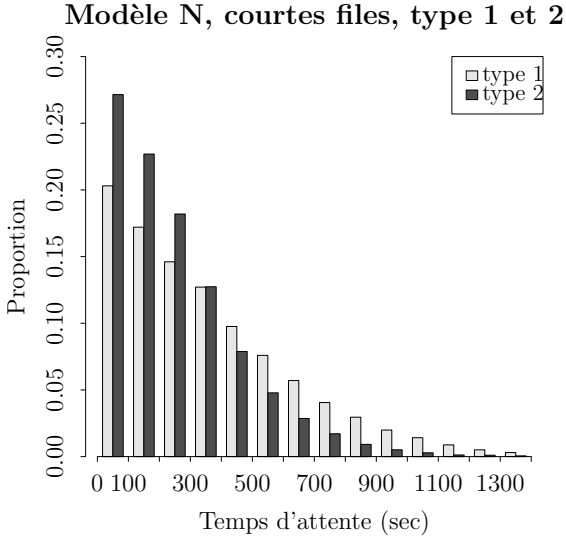

Figure 4.1 : Modèle N avec courtes files : Distribution du temps d'attente des clients qui ont attendu et reçu le service, pour chaque type.

Type d'appel	LES	RS	RS(t,q)	SK	SK(t,q)	ANN	ANN(t,q)
1	0.871	0.603	0.612	0.590	0.602	**0.572**	0.581
2	0.850	0.597	0.600	0.591	0.603	**0.560**	0.571

Tableau 4.11 : Le RRASE pour le N. modèle avec de courtes files.

Le tableau 4.11 rapporte les RRASEs. Il montre que nos prédicteurs RS, SK et ANN et leurs variantes ont une RRASE qui est environ 25% inférieure à celle du prédicteur LES, pour tous les types d'appels. Le paramètre $\boldsymbol{r}$ (la longueur de la file d'attente secondaire) dans l'entrée $\boldsymbol{x}$ a peu d'effet sur la précision de RS, SK et ANN. ANN est plus performant que SK. Ce dernier est plus aussi un peu plus performant que RS. Cette remarque est aussi faite sur leurs variantes. La

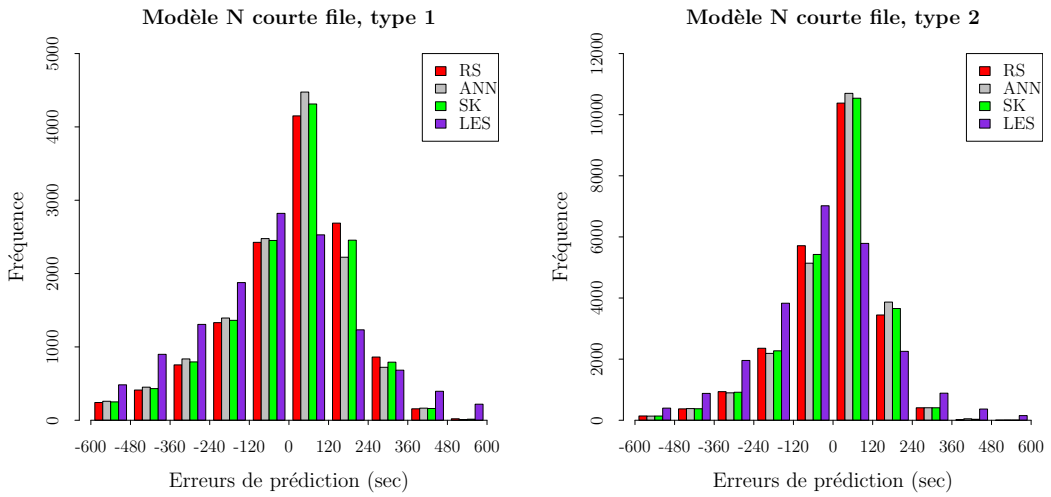

Figure 4.2 : Modèle N avec courtes files : Distribution des erreurs de prédiction (délai estimé moins délai réel) pour le type 1 et type 2.

figure 5.6 donne un histogramme de l'erreur de prédiction pour chaque méthode (à l'exclusion des variantes). Nous voyons que le prédicteur LES a une distribution d'erreur beaucoup plus variable (de grandes erreurs sont plus fréquentes), tandis que RS, SK et ANN ont des distributions d'erreur très similaires, pour les deux types d'appels.

4.3.4.2 Un modèle N avec de longues files d'attente

Notre deuxième exemple de modèle N a de longues files d'attente. Nous avons pris de plus grands taux d'arrivée et des temps de patience que dans le cas précédent, tout en gardant le staffing presque inchangé. Pour les appels de type 1, les taux d'arrivée sont $\lambda_1 = (25, 34, 43, 48, 51, 57, 42, 34, 22, 18)$ par heure, le temps de service moyen est $\mu_1^{-1} = 21$ minutes, et la moyenne du temps de patience est $\nu_1^{-1} = 46.7$ minutes. Pour le type 2, les taux d'arrivée sont $\lambda_2 = (26, 40, 47, 59, 68, 59, 48, 43, 39, 29)$ par heure, le temps de service moyen est $\mu_2^{-1} = 11$ minutes, et la moyenne du temps de patience est $\nu_2^{-1} = 30$ minutes. Les vecteurs de staffing sont $s_1 = (4, 6, 9, 10, 9, 9, 9, 8, 5, 5)$ et $s_2 = (4, 7, 9, 10, 9, 8, 7, 8, 6, 5)$.

Les mesures de performance agrégées de toutes les périodes sont les suivantes.

Pour l'appel de type 1, nous trouvons une moyenne de 9.7 clients dans la file d'attente, une proportion de délais de 94%, un ratio d'abandon de 33%, un temps d'attente moyen de 938 secondes pour tous les appels, et un temps d'attente moyen de 1151 secondes pour les appels qui sont entrés dans la file d'attente et ont été servis. Pour le type 2, ces mesures sont 5.5 clients, 97% mis en attente, 23% des abandons, 426 secondes et 465 secondes, respectivement. Dans cet exemple, 88% des appels servis de type 1 ont été répondus par le groupe 1 et les 12% restants ont été répondus par le groupe 2. La figure 4.3 montre que les distributions du temps d'attente pour les clients qui ont attendu et servi sont très différentes entre les types d'appels 1 et 2.

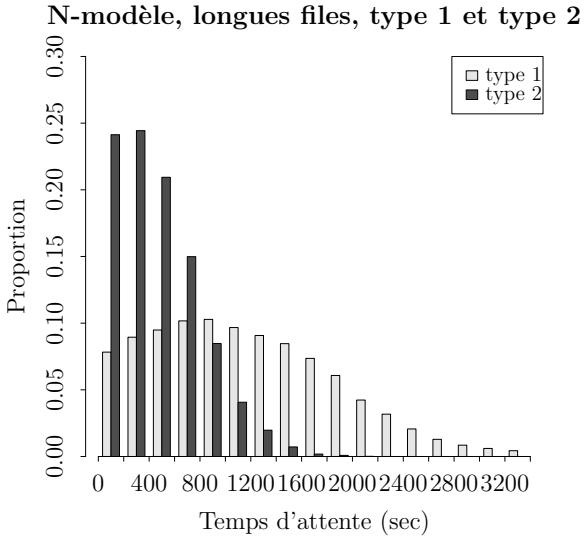

Figure 4.3 : Modèle N avec longues files : Distribution du temps d'attente pour les clients qui ont attendu et servi.

	Types d'appels						
Type d'appel	LES	RS	RS(t,q)	SK	SK(t,q)	ANN	ANN(t,q)
1	0.499	0.364	0.369	0.360	0.370	**0.341**	0.350
2	0.629	0.443	0.446	0.445	0.458	**0.424**	0.435
Global	0.581	0.418	0.422	0.408	0.418	**0.393**	0.407

Tableau 4.12 : Les RRASEs pour le modèle N avec longues files.

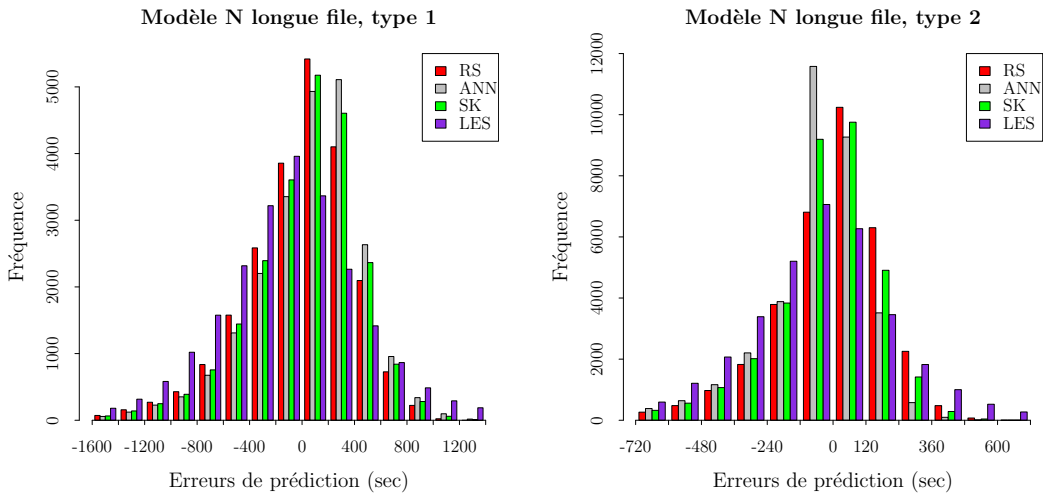

Figure 4.4 : Modèle N avec longues files : Distribution des erreurs de prédiction (délai estimé moins délai réel) pour les types d'appels 1 et 2.

En regardant les RRASEs dans le tableau 4.12, nous constatons que la comparaison entre les prédicteurs est très similaire à ce que nous avons vu dans l'exemple précédent avec les courtes files d'attente. LES est beaucoup moins performant que toutes nos nouvelles méthodes ; son RRASE est plus grand de 13% à 19%. Encore une fois, ANN est un peu mieux que SK et qui à son tour est un peu mieux que RS. L'ajout du paramètre $\boldsymbol{r}$ à l'entrée $\boldsymbol{x}$ améliore légèrement la précision des nouveaux prédicteurs. La Figure 5.7 illustre également une plus grande variance dans l'erreur de prédiction avec LES.

4.3.5 Expériences avec un grand centre d'appels basé sur des données réelles (HQ)

Nous considérons un grand exemple inspiré par les données d'un sous-ensemble d'appels et d'agents d'un véritable centre d'appels d'un fournisseur de services publics au Québec. Le centre fonctionne de 8 heures à 18 heures dans une journée. Ces heures d'ouverture sont divisées en 40 périodes de 15 minutes. Le centre d'appels global gère 96 types d'appels avec 375 groupes d'agents, mais nous avons sélectionné les 6 types d'appels ayant les plus gros volumes et 8 groupes d'agents qui

peuvent les servir, comme dans Chan et al. (2014).

Les ensembles de compétence des 8 groupes sont $S_1 = \{1, 3, 4, 5\}$, $S_2 = \{1, 2\}$, $S_3 = \{3, 5\}$, $S_4 = \{3, 5, 6\}$, $S_5 = \{1, 3, 5\}$, $S_6 = \{1, 2, 3, 5\}$, $S_7 = \{3, 5, 6\}$, and $S_8 = \{1, 3, 5, 6\}$. Les arrivées sont des processus de Poisson avec des taux constants $\lambda_{j,p}$ dans la période p pour chaque type j. Le vecteur de la moyenne agrégée des taux d'arrivées de la journée pour les 6 types d'appels est $(35.5, 6.0, 98, 6.5, 29, 3.5)$. Les temps de patience sont exponentiels avec une moyenne (pour les six types d'appels) : $(52, 36, 41, 51, 41, 15)$. Les temps de service sont de loi log-normale, avec des paramètres différents pour chaque paire de groupe d'agent g et type d'appel j qui peut être servi par ce groupe. Les moyennes estimées varient de 5.14 à 11.3 minutes, et les écarts-types varient de 5.88 à 22.0 minutes.

Les paramètres ont été légèrement modifiés par rapport à ceux du véritable centre pour des raisons de confidentialité. Cet exemple a la particularité que le taux d'arrivée et le staffing changent considérablement au cours de la journée.

Nous avons simulé 100 jours indépendants, comme pour les autres exemples. Le tableau 4.13 montre les mesures de performances globales pour les six types d'appels. Notez que la longueur moyenne de file d'attente varie considérablement selon les types d'appels (d'environ 3 à 120). Le type d'appel 6, dont la longueur de la file et le temps d'attente moyen sont plus petits, en fait a une grande priorité pour tous les groupes qui peuvent le servir. Le type d'appel 3 a une priorité plus faible et le plus grand volume.

	Types d'appels					
Performance	1	2	3	4	5	6
PD(%)	98.1	99.5	98.6	97.9	97.8	99.2
PA(%)	34.2	37	43	34.4	39.2	13.6
AQS	41.3	6.65	120	6.95	38.2	2.96
AWT (Sec.)	1037	1078	1610	1053	1125	208

Tableau 4.13 : Mesures de performances moyennes pour le grand exemple.

Le tableau 4.14 montre les RRASEs×100 pour les six types d'appels. Nous constatons toujours que les prédicteurs ANN, SK, RS sont largement plus précis

que LES pour les six types d'appels. La différence de précision varie entre 10% et 20% pour les six types d'appels. ANN est toujours le prédicteur le plus performant sauf pour le type d'appel 6. SK arrive en seconde position et il est suivi par RS. Comme dans les exemples précédents, nous observons que la différence de précision entre RS, SK et ANN n'est pas trop grande. Elle varie entre 1 et 4% pour tous les 6 types d'appels.

Nous constatons que le type d'appels 6, qui a un faible taux d'arrivée, une courte file d'attente, un petit temps d'attente moyen et qui est prioritaire pour tous les groupes d'agents qui possèdent la compétence pour le traiter, a un RRASE beaucoup plus élevé que les autres. Le RRASE de ce type dépasse 61% pour tous les prédicteurs et est environ 4 fois plus grand que celui des autres types. Une autre remarque : il est l'unique cas dans l'étude des systèmes multi-compétences ou RS est plus précis que ANN. Les appels de type 6 sont rares et par conséquent nous aurons peu de données pour ce type à la fin de la simulation du modèle. Nous pensons que les prédicteurs n'ont pas pu apprendre les bons paramètres avec les données disponibles. Les prédicteurs qui utilisent l'apprentissage machine ont besoin de beaucoup de données pour apprendre les bons paramètres de la fonction de prédictions. ANN est celui qui a plus besoin de données pour bien performer. C'est la raison pour laquelle il est moins performant que RS pour ce type d'appel. Nous allons y revenir dans la section 4.5 qui est consacrée à l'étude de la robustesse des nouveaux prédicteurs.

Le temps d'entraînement des modèles de prédictions est grand pour les types 1, 2, 3, 4 et 5 avec cet exemple. Il est en moyenne de 5 minutes pour RS, de 12 heures de pour ANN, et de 24 heures pour SK. Le temps d'entraînement du modèle augmente toujours quand la taille de l'échantillon d'entraînement ou la taille du vecteur qui définit l'état du système augmente.

	Types d'appels					
	1	2	3	4	5	6
LES	24.6	35.6	20.3	41.3	26.1	94.5
RS	8.90	12.9	11.4	15.9	18.9	62.7
SK	8.50	11.6	10.1	14.0	16.5	**61.5**
ANN	**7.50**	**10.1**	**8.20**	**12.2**	**14.8**	63.7

Tableau 4.14 : RRASEs pour les 6 types d'appels du grand exemple.

4.4 Impact de l'ajout de la période et du nombre d'agents des groupes dans la définition de l'état du système

Dans cette section, nous étudions l'impact de l'ajout de nouvelles informations dans la définition de l'état du système sur la précision des prédictions. Pour chaque type d'appel k, l'état du système était défini jusqu'à présent par $\boldsymbol{x} = (t, q, \boldsymbol{r})$ où t le délai d'attente du dernier client de type k entré en service, q la longueur de la file d'attente k, et $\boldsymbol{r}$ est un vecteur qui contient la taille de chaque file d'attente $j \neq k$ tel qu'il y ait au moins un agent avec les deux compétences k et j.

Nous considérons, en plus de ces trois informations, deux nouvelles informations dans la définition de l'état du système. Le premier est la période p d'arrivée de l'appel. Le second est un vecteur $\boldsymbol{s} = (s_1, \ldots, s_G)$ qui représente le staffing des groupes d'agents à la période d'arrivée de l'appel (pour tout $g \in \{1, \ldots, G\}$, s_g est le staffing du groupe g à la période d'arrivée de l'appel de type k).

Dans les centres d'appels, les taux d'arrivées des appels, le staffing des groupes d'agents varient souvent d'une période à une autre de la journée. Le temps d'attente d'un nouvel appel qui entre dans une file d'attente pourrait bien dépendre de la période p d'arrivée et du staffing $\boldsymbol{s}$ des groupes qui servent les appels.

Nous définissons un nouvel état du système par $\boldsymbol{x} = (t, q, \boldsymbol{r}, p, \boldsymbol{s})$. En ajoutant p et $\boldsymbol{s}$ aux informations qui décrivent l'état du système, nos prédicteurs deviennent dépendants du temps. Les nouveaux prédicteurs RS, SK et ANN qui prennent en entrée $\boldsymbol{x} = (t, q, \boldsymbol{r}, p, \boldsymbol{s})$ seront notés respectivement par $\mathrm{RS}(t, q, \boldsymbol{r}, p, \boldsymbol{s})$, $\mathrm{SK}(t, q, \boldsymbol{r}, p, \boldsymbol{s})$, et $\mathrm{ANN}(t, q, \boldsymbol{r}, p, \boldsymbol{s})$.

Nous allons maintenant comparer la précision des prédictions en utilisant ce nou-

veau vecteur d'entrée dans les trois exemples de centres d'appels multi-compétences qui sont déjà étudiés à la section 4.3. Le premier et le deuxième sont les modèles N avec courtes files d'attente et longues file d'attente. Le troisième exemple celui basé sur des données réelles.

4.4.1 Modèle N avec longues files

Le tableau 4.15 donne les RRASEs des prédicteurs pour le modèle N avec longues files étudié à la section 4.3.4.2 pour les états $\boldsymbol{x} = (t, q, \boldsymbol{r})$ et $\boldsymbol{x} = (t, q, \boldsymbol{r}, p, \boldsymbol{s})$. Pour chaque type prédicteur, l'ajout de p et $\boldsymbol{s}$ dans la définition de l'état du système entraînent la réduction du RRASE. Cette réduction est environ de 2% pour les deux types d'appels avec les différents prédicteurs. Nous observons toujours que ANN$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$ est plus précis que SK$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$ qui a son tour est plus précis que RS$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$.

Cependant, nous notons que le temps d'entraînement des modèles de prédiction est plus beaucoup plus grand pour $\boldsymbol{x} = (t, q, \boldsymbol{r}, p, \boldsymbol{s})$ que pour $\boldsymbol{x} = (t, q, \boldsymbol{r})$. Les temps d'entraînement qui étaient en moyenne de 2, 30, et 60 minutes pour RS, ANN et SK respectivement sont maintenant en moyenne de 4, 70, 190 minutes pour RS$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$, ANN$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$, SK$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$.

Type	LES	RS	RS$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$	SK	SK$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$	ANN	ANN$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$
1	0.499	0.364	0.344	0.360	0.355	0.341	**0.321**
2	0.629	0.443	0.418	0.445	0.436	0.424	**0.411**

Tableau 4.15 : Les RRASEs pour le modèle N avec longues files.

4.4.2 Le modèle N avec courtes files

Le tableau 4.16 rapporte les RRASEs des prédicteurs pour le modèle N courtes files étudié à la section 4.3.4.1 pour $\boldsymbol{x} = (t, q, \boldsymbol{r}, p, \boldsymbol{s})$. Comme dans l'exemple avec longue file, nous observons une réduction du RRASE avec les différents prédicteurs. Cette diminution est d'environ 2.5% pour les deux types d'appels. Dans cet exemple aussi, ANN$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$ est toujours le prédicteur le plus précis. Il est

suivi de SK$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$. RS$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$ reste toujours le prédicteur le moins précis. Dans cet exemple aussi nous constatons que le temps d'entraînement du modèle en moyenne a doublé pour RS et ANN tandis que pour SK, il a plus que triplé.

Type	LES	RS	RS$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$	SK	SK$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$	ANN	ANN$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$
1	0.871	0.603	0.589	0.590	0.572	0.572	**0.542**
2	0.850	0.597	0.573	0.591	0.568	0.560	**0.536**

Tableau 4.16 : Le RRASE pour le N. modèle avec de courtes files.

4.4.3 Exemple avec un grand centre d'appels basé sur des données réelles

Le tableau 4.17 montre les RRASEs des prédicteurs pour les six types d'appels. Comme dans les deux exemples précédents, nous constatons que l'ajout d'informations (la période d'arrivée et du staffing des groupes) dans la définition de l'état du système a augmenté la précision des prédicteurs. La diminution du RRASE varie entre 1 et 5% pour les six types d'appels. Cependant, nous notons une augmentation considérable du temps d'entraînement des modèles de prédiction pour RS et SK. En moyenne, il est de 48 heures pour SK et de 24 heures pour ANN. Pour RS le temps d'entraînement est toujours de quelques minutes. Il varie entre 3 et 5 minutes pour les six types d'appels.

	Types d'appels					
	1	2	3	4	5	6
LES	24.6	35.6	20.3	41.3	26.1	94.5
RS	8.90	12.9	11.4	15.9	18.9	62.7
RS$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$	7.20	10.2	9.8	14.3	16.7	60.1
SK	8.50	11.6	10.1	14.0	16.5	61.5
SK$(t, q, \boldsymbol{r}, p, \boldsymbol{s})$	7.80	9.92	8.86	12.2	14.8	**59.1**
ANN	7.50	10.1	8.20	12.2	14.8	63.7
ANN $(t, q, \boldsymbol{r}, p, \boldsymbol{s})$	**5.70**	**8.5**	**7.12**	**10.5**	**12.3**	62.4

Tableau 4.17 : RRASEs pour les 6 types d'appel de l'exemple basé sur des données réelles.

Dans tous les exemples considérés, nous constatons l'ajout des informations

dans la définition de l'état du système entraîne une amélioration de la précision des prédictions. La réduction du RRASE varie entre 1 et 5 %. Le seul inconvénient est que l'entraînement du modèle peut prendre beaucoup de temps pour ANN et SK. Nous avons constaté plus que la taille du vecteur $\boldsymbol{x}$ est grande, plus le temps nécessaire pour apprendre le bon vecteur de paramètres θ augmente.

ANN est le prédicteur plus précis, mais il demande le plus souvent plusieurs essais pour trouver le bon réseau (combien de couches cachées, combien de neurones par couche). Nous n'avons pas une méthode rapide pour déterminer les bons paramètres du réseau. Le temps nécessaire pour trouver les bons paramètres peut souvent dépasser plusieurs heures surtout si la taille du vecteur $\boldsymbol{x}$ est grande. Si la taille des données d'entraînement est petite alors il est préférable d'utiliser RS.

SK est la plupart du temps moins bon que ANN et en plus le temps nécessaire pour entraîner le modèle est beaucoup plus grand que celui de ANN. Il est en général plus précis que RS, mais la différence de précision varie le plus souvent entre 1 et 3%.

RS est souvent le prédicteur le moins précis, mais son temps d'entraînement est largement plus petit que ceux de ANN et SK. Il a l'avantage d'être rapide pour trouver le bon vecteur de paramètres θ. En général quelques minutes suffirent pour le trouver. Il est aussi le prédicteur le plus précis si la taille des données d'entraînement est petite.

Le choix d'un prédicteur par un gestionnaire dépend du niveau de précision des prédictions voulu, du temps dont on dispose pour entraîner le modèle et de la quantité de données disponibles.

4.5 La Robustesse des prédicteurs

Nos prédicteurs sont développés (optimisés) en utilisant des données (les données d'entraînement) obtenues à partir d'une simulation du modèle. Jusqu'à présent les données de *test*, qui sont utilisées pour comparer le RRASE des différents prédicteurs, sont aussi générées par une autre simulation indépendante du modèle tout

en gardant toujours les mêmes paramètres du modèle. Dans la vraie vie, nous savons que les taux d'arrivée des types d'appels prédits sont toujours différents de ceux qui sont réellement observés. Nous observons soit des taux plus petits ou soit des taux plus élevés. Il est bon de voir la précision de nos prédicteurs face à ces situations.

Nous savons que les vraies données dans les centres d'appels sont toujours collectées sur une longue période de plusieurs journées et nous savons aussi que les taux d'arrivées des appels à travers les journées sont différents. Il est intéressant d'observer la précision des prédicteurs dans le cas où les données d'entraînement du modèle sont obtenues sur des journées différentes.

À la section 4.3, nous avons étudié plusieurs exemples de modèles de complexité différente. Dans tous les cas, nous avons enregistré beaucoup de données pour tous les types d'appels et les algorithmes d'apprentissage machine sont parvenus à apprendre les bons paramètres de la fonction de prédictions. Il est bien connu que ces algorithmes sont en général moins performants s'il n'y a pas assez de données. Il est intéressant d'observer et de comparer la précision des prédicteurs pour des types qui reçoivent très peu d'appels et pour lesquels nous avons enregistré peu de données.

Dans cette sous-section, nous allons en premier observer le comportement des prédicteurs dans le cas où les données de *test* sont générées en utilisant des taux d'arrivée légèrement différents de ceux des données d'entraînement. Ensuite, nous allons étudier la précision des prédicteurs dans le cas où les données d'entraînement sont obtenues sur plusieurs journées différentes.

Pour obtenir les nouvelles données de test, nous simulons un modèle avec des taux d'arrivée légèrement modifiés. Pour chaque type d'appel k, le nouveau taux d'arrivée à la période p, noté $\tilde{\lambda}_{k,p}$, sera modifié comme suit :

$$\tilde{\lambda}_{k,p} = \lambda_{k,p} + \sigma \cdot \lambda_{k,p}, \tag{4.3}$$

où σ représente le pourcentage de diminution ($\sigma < 0$) ou d'augmentation ($\sigma > 0$)

sur le taux initial $\lambda_{k,p}$ du type k à la période p. Le taux initial $\lambda_{k,p}$ est le taux d'arrivée du type k à la période p utilisé pour générer les données d'entraînement. Dans nos exemples numériques, nous diminuons ou augmentons le même pourcentage sur le taux de tous les types simultanément.

Pour obtenir des données d'entraînement collectées sur plusieurs journées avec des taux d'arrivée différents, nous procédons comme suit. Soit λ_k le taux d'arrivée du type d'appel k à la première journée. Pour toute autre journée, le taux d'arrivée du type k est donné par $\mathrm{B}\lambda_k$ où B est le *busyness factor* de la journée qui est une variable aléatoire Gamma de moyenne $m = 1$ et variance v.

4.5.1 Variation du taux d'arrivée avec le modèle M/M/s+M

Nous considérons le même exemple étudié à la section 4.3.1. Nous rappelons qu'avec cet exemple les paramètres du modèle étaient les suivants. Le taux d'arrivée est $\lambda = 50$ appels par heure, les durées de service ont une moyenne $\mu^{-1} = 2$, les abandons ont une moyenne $\nu^{-1} = 0.5$ heure, et le nombre d'agents qui répondent aux appels est $s = 17$. Pour entraîner RS, SK et ANN, nous générons les données d'entraînement par une simulation du modèle avec les paramètres énumérés ci-dessus. Dans cet exemple, nous simulons une seule et longue période pour un seul type d'appel avec un seul groupe. En plus de la variation du taux d'arrivée, nous considérons un autre cas où les données d'*entraînement* du prédicteur sont collectées sur plusieurs journées différentes.

Données de test avec des taux d'arrivée différents

Pour chaque $\sigma \in \{-0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, nous avons généré des données de test, en utilisant un nouveau taux d'arrivée déterminé par l'équation (4.3). Le tableau 4.18 donne les moyennes globales des mesures de performances du centre d'appels avec chacun des taux d'arrivée utilisés. Les mesures de performances observées dans ce tableau sont la longueur de la file d'attente moyenne, la proportion de délais, la proportion d'abandons, et le temps d'attente moyen. Nous

observons que les performances des modèles simulés pour générer les données de *test* sont largement différentes de celles du modèle d'entraînement.

	σ							
Performances	-20%	-10%	0	+10%	+20%	+30%	+40%	+50%
AQL(clients)	4.41	6.67	9.04	11.51	14.00	16.49	18.99	21.50
PD(%)	84.33	93.64	97.77	99.30	99.80	99.94	99.98	99.99
PA(%)	31.18	29.54	36.18	41.86	46.68	50.77	54.28	57.33
AWT(seconds)	397	532	651	753	839	913	976	1031

Tableau 4.18 : Performances du modèle M/M/s+M avec les variations du taux d'arrivée

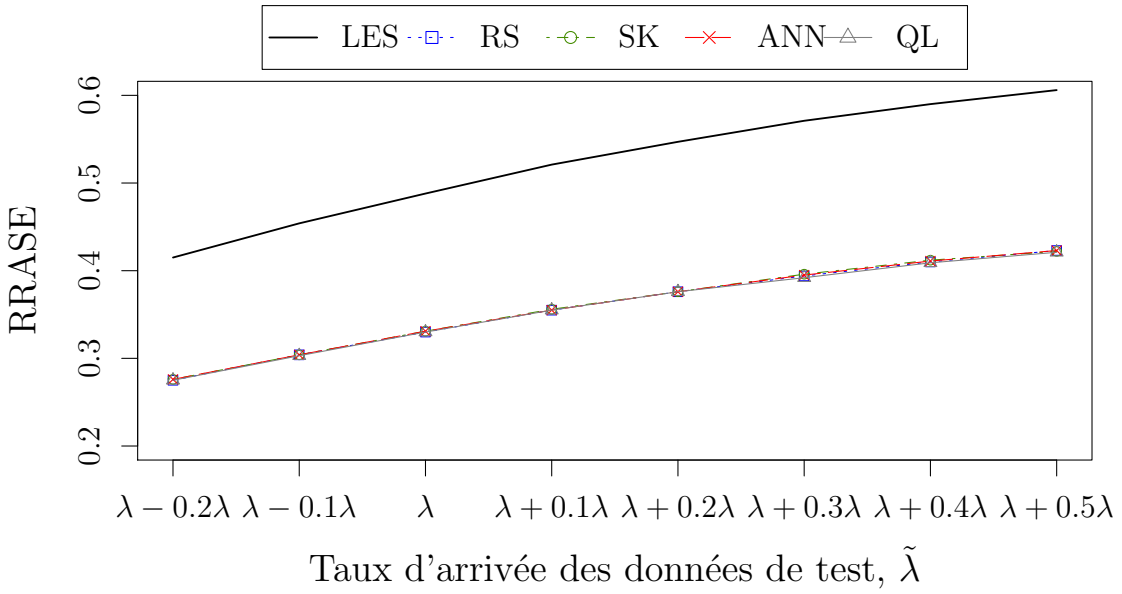

Figure 4.5 : RRASE des prédicteurs pour modèle M/M/s+M en fonction du taux d'arrivée des données de test.

Le tableau 4.19 rapporte les RRASEs observés avec les prédicteurs dans chacune des cas. Nous constatons que la variation du taux d'arrivée (augmentation ou diminution) n'affecte pas beaucoup la précision des prédicteurs RS, SK et ANN. Les nouveaux prédicteurs ont une précision toujours très similaire à celle du prédicteur optimal QL dans les différents cas. Ceci s'explique par le fait que la fonction de

prédiction dans les systèmes à file unique est indépendante du taux d'arrivée. Nos observons aussi que la différence de précision entre LES et les nouveaux prédicteurs reste sensiblement inchangée.

	σ							
Prédicteur	-20%	-10%	0	+10%	+20%	+30%	+40%	+50%
LES	0.415	0.454	0.488	0.521	0.547	0.571	0.590	0.606
RS	0.275	0.304	0.330	0.355	0.376	0.394	0.410	0.423
ANN	0.276	0.304	0.331	0.355	0.376	0.395	0.411	0.423
QL	0.275	0.303	0.330	0.355	0.376	0.392	0.409	0.421
SK	0.276	0.304	0.331	0.356	0.376	0.396	0.412	0.422

Tableau 4.19 : RRASE M/M/s+M avec variation du taux d'arrivée

Le RRASE des prédicteurs est fonction du taux d'arrivée des données de test (même pour le prédicteur optimal QL). Elle augmente si le taux d'arrivée au système augmente et diminue si le taux diminue. La figure 4.5 affiche le RRASE des prédicteurs en fonction du taux d'arrivée $\tilde{\lambda}$.

Données d'entraînement avec des taux d'arrivée différents

Nous considérons toujours le centre d'appels modélisé par une file d'attente M/M/s+M. Les paramètres du centre sont toujours les mêmes sauf pour le taux d'arrivée λ. Dans cet exemple $\lambda = 50 \times B$ où B est une variable aléatoire qui représente le facteur d'achalandage ou "*busyness factor*" de la journée. B suit une distribution gamma de paramètres $\alpha = 50$ et $\beta = 50$. La moyenne des durées de service $\mu^{-1} = 2$, la moyenne des temps de patience est $\nu^{-1} = 0.5$ heure, et le nombre d'agents $s = 17$.

Nous avons généré par simulation les données de 5 journées indépendantes. Par la suite, nous avons utilisé les données des quatre premières journées comme données d'entraînement. Les données de la cinquième journée sont utilisées pour calculer le RRASE des prédicteurs.

Le tableau 4.20 donne les RRASEs des différents prédicteurs. Nous remarquons que les prédicteurs QL, ANN, RS et SK ont toujours une précision très similaire.

Prédicteurs	LES	RS	SK	ANN	QL
RRASE	0.607	0.419	0.413	0.419	**0.408**

Tableau 4.20 : RRASE des prédicteurs pour un centre d'appels M/M/s+M avec "busyness factor".

La différence de précision entre LES et les autres prédicteurs est sensiblement la même que dans le cas initial. Cet exemple montre que les nouveaux prédicteurs sont robustes même si les données d'entraînement sont collectées à partir de plusieurs journées. Ce résultat est encourageant, car dans la vraie vie les données d'entraînement sont toujours obtenues sur plusieurs journées distinctes.

4.5.2 Variation du taux d'arrivée avec le modèle N

Nous considérons maintenant le modèle N étudié à la section 4.3.4.2.

Données de test avec des taux d'arrivée différents

Pour chaque $\sigma \in \{-0.2, -0.1, 0, 0.1, 0.2, 0.3\}$, nous générons de nouvelles données de test en utilisant des taux d'arrivées déterminés par l'équation (4.3). Le tableau 4.21 donne le RRASE des prédicteurs avec les différentes données de test. Comme dans l'exemple précédent, nous constatons que le changement du taux d'arrivée ne réduit pas beaucoup la précision des prédicteurs RS, SK et ANN. En effet, les RRASEs obtenus avec ces données de test sont sensiblement égaux aux RRASEs des prédicteurs avec des données d'entraînement de taux d'arrivée égal au taux d'arrivée des données de test. La différence de précision entre les nouveaux prédicteurs et LES est aussi sensiblement égale à leur différence de précision dans le cas initial. L'ordre de précision des prédicteurs est toujours le même. ANN est plus précis que SK qui a son tour est plus précis que RS. Cela montre que les informations que nous avons choisies pour décrire l'état du système sont suffisantes pour apprendre la fonction de prédiction.

La figure 4.6 et la 4.7 affichent les prédicteurs en fonction du taux des données de test pour les types 1 et 2, respectivement. Elles montrent que dans cet exemple

aussi que le RRASE est fonction du taux d'arrivée des données de test.

Predicteurs	-20%				-10%			
	LES	RS	SK	ANN	LES	RS	SK	ANN
T1	0.372	0.266	0.260	**0.250**	0.430	0.309	0.306	**0.301**
T2	0.566	0.398	0.383	**0.369**	0.597	0.428	0.402	**0.409**

Prédicteurs	0%				10%			
	LES	RS	SK	ANN	LES	RS	SK	ANN
T1	0.499	0.364	0.360	**0.351**	0.543	0.385	0.375	**0.368**
T2	0.629	0.459	0.445	**0.434**	0.674	0.487	0.474	**0.464**

	+20%				+30%			
	LES	RS	SK	ANN	LES	RS	SK	ANN
T1	0.579	0.423	0.416	**0.395**	0.609	0.463	0.458	**0.425**
T2	0.703	0.502	0.503	**0.490**	0.739	0.537	0.519	**0.521**

Tableau 4.21 : RRASE du modèle N avec variation du taux d'arrivée

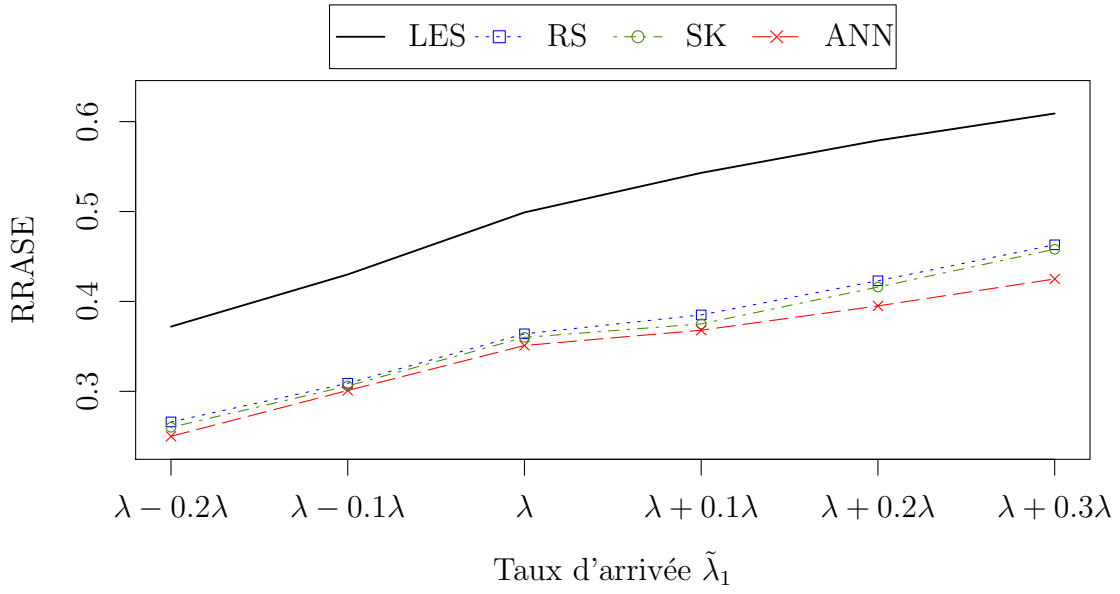

Figure 4.6 : RRASE des prédicteurs pour le modèle N en fonction du taux d'arrivée $\tilde{\lambda}_1$, type 1.

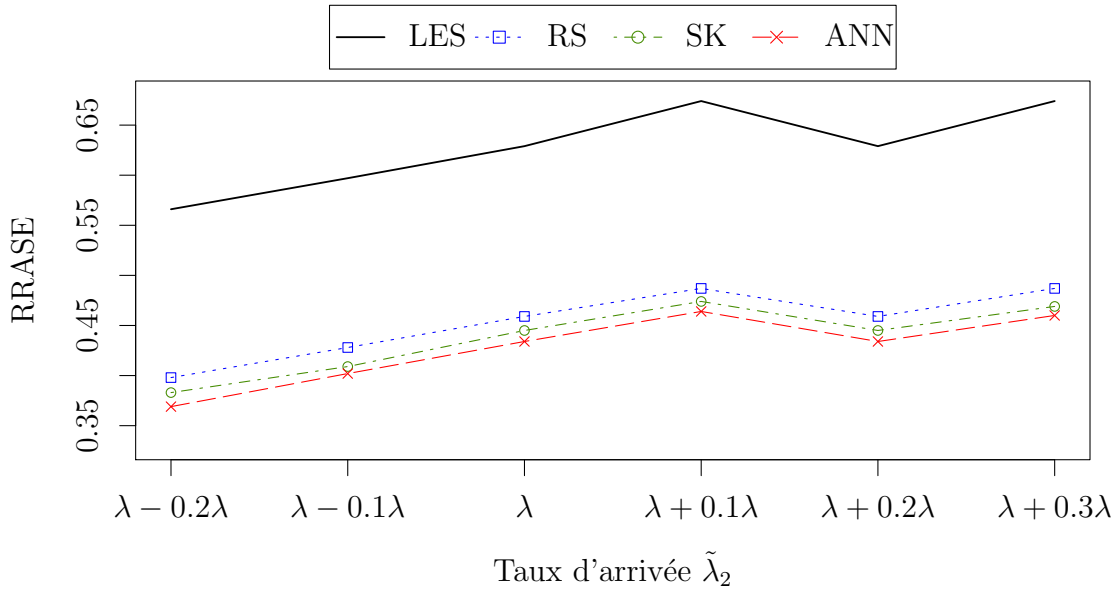

Figure 4.7 : RRASE des prédicteurs pour le modèle N en fonction du taux d'arrivée $\tilde{\lambda}_2$, type 2.

Données d'entraînement avec des taux d'arrivée différents

Nous avons simulé 100 journées. Pour chacune des journées le nouveau taux d'arrivée du type 1, $\tilde{\lambda}_1$, et le nouveau taux d'arrivée du type 2, $\tilde{\lambda}_2$, sont obtenus par $\tilde{\lambda}_1 = B\lambda_1$ et $\tilde{\lambda}_2 = B\lambda_1$ où B le *busyness factor* de la journée. Il suit une distribution Gamma de paramètres $\alpha = 50$ et $\beta = 50$. Nous avons utilisé les données des 80 premières journées pour entraîner les modèles et les données des 20 journées restantes pour évaluer les performances des prédicteurs. Le tableau rapporte les RRASEs des différents prédicteurs pour les deux types d'appels. Comme dans le cas de la file simple, nous observons le fait d'entraîner les modèles sur des données provenant de plusieurs journées différentes ne diminue pas beaucoup la précision des prédicteurs. En effet, si nous comparons les résultats des prédicteurs avec ceux obtenus à la section 4.3.4.2 qui étudie le même exemple avec des données d'entraînement et des données de test générées avec λ_1 et λ_2, nous observons une augmentons des RRASEs d'environ 2%. L'ordre de précision des prédicteurs est

toujours le même. ANN est plus précis que SK qui à son tour est plus précis que RS.

Type	LES	RS	SK	ANN
1	0.525	0.382	0.384	**0.360**
2	0.646	0.464	0.463	**0.442**

Tableau 4.22 : Les RRASEs pour le modèle N avec données d'entraînement et de test collectées sur plusieurs journées.

4.5.3 La précision des prédicteurs RS, ANN et SK pour les types appels rares

Nous allons maintenant observer la robustesse des prédicteurs pour les types d'appels avec un faible taux d'arrivée. Nous considérons un modèle W avec 3 types d'appels et deux groupes d'agents, où les groupes ont les ensembles de compétences $\mathcal{S}_1 = \{1, 2\}$ et $\mathcal{S}_2 = \{2, 3\}$. Le groupe 1 traite les appels de type 1 et de type 2 et le groupe 2 traite les appels de type 2 et type 3.

Les processus d'arrivée sont des processus de Poisson de taux constant durant la journée. Toutes les durées de service et les temps de patience sont exponentiels et indépendants. Nous utilisons une politique de *routage par priorité* qui fonctionne comme suit. Pour un appel de type 2, le routeur va d'abord essayer de l'affecter à un agent libre du groupe 1. S'il n'y a pas d'agent libre, alors le routeur va essayer de l'attribuer à un agent libre du groupe 2. Les agents du groupe 1 accordent la même priorité aux appels de type 1 et aux appels de type 2. Si un agent du groupe 1 devient libre et qu'il y a un appel 1 et un appel 2 en attente, alors la priorité sera accordée à l'appel qui a enregistré la plus longue attente, et si les deux appels ont observé le même temps, le routeur choisit un appel au hasard. Les agents du groupe 2 donnent toujours la priorité aux appels de type 2, même si certains appels de type 3 ont attendu plus longtemps. Ainsi les appels type 3 sont servis si seulement il n'y a pas d'appel de type 2 en attente. Les appels du même type sont de *premier arrivé, premier servi*, mais les appels de différents types peuvent être servis dans des ordres différents.

Nous supposons que les appels de type 3, qui ne sont pas prioritaires pour le groupe 2, ont un taux d'arrivée faible, et les appels type 1 et type 2 ont des taux d'arrivée élevés. Après une simulation du modèle, nous pourrons enregistrer beaucoup de données pour les types 1 et 2, par contre pour le type 3, nous allons enregistrer peu de données. Ce qui nous intéresse le plus dans cet exemple est le comportement des prédicteurs RS, SK et ANN dans la prédiction des temps d'attente pour le type 3.

Le vecteur des taux d'arrivées pour les trois types d'appels est $(\lambda_1, \lambda_2, \lambda_3) = (20, 10, 1)$, le vecteur des taux de service $(\mu_1, \mu_2, \mu_3) = (2, 2, 1)$ et celui des temps moyens de patience $(\nu_1, \nu_2, \nu_3) = (0.5, 0.2, 0.3)$.

Nous avons simulé 100 jours indépendants, comme pour les autres exemples. Le tableau 4.23 montre les mesures de performances moyennes pour les 3 types d'appels. Notez que la longueur moyenne de file d'attente varie selon les types d'appels (d'environ 1 à 10). Le type d'appel 3, dont la longueur de la file est petite et le temps d'attente moyen est grand, en fait a une faible priorité pour le groupe qui peut le servir.

	Types d'appels		
Performances	1	2	3
PD(%)	94.1	87.5	88.7
PA(%)	25.5	18	37
AQS	9.5	3.4	1.1
AWT (min)	30	21.3	104.4

Tableau 4.23 : Mesures de performances moyennes pour le modèle W..

Le tableau 4.24 montre les RRASEs$\times$100 pour les 3 types d'appels. Nous constatons que pour les types 1 et 2 que ANN donne toujours les meilleurs résultats. Il est suivi par RS et SK. LES est toujours le prédicteur le moins précis. Par contre pour le type d'appel 3, nous observons que les prédicteurs RS, SK et ANN sont beaucoup moins précis avec des RRASE beaucoup plus grand que ceux des deux autres types. Ils ont des RRASE qui sont tous supérieurs à 60%. En plus de cela, nous constatons que RS est plus précis que ANN, qui a son tour est un peu plus précis

que SK. Cependant, nous notons que les prédicteurs ANN, SK, RS sont toujours plus précis que LES.

Type d'appel	LES	RS	SK	ANN
1	44.9	30.4	31.6	**28.3**
2	65.0	39.1	40.5	**35.7**
3	115.8	**60.6**	67.1	66.9

Tableau 4.24 : Les RRASEs pour le modèle W.

Les mauvaises performances des prédicteurs qui utilisent l'apprentissage machine pour le type d'appel 3 sont dû à la quantité de données disponibles pour entraîner les modèles. Les données ne sont pas suffisantes pour apprendre les bons paramètres de la fonction de prédiction. Comme dans l'exemple basé sur des données réelles étudié à la section 4.3.5 de ce chapitre, nous observons que dans une telle situation, le prédicteur RS est préférable par rapport aux prédicteurs SK et ANN. Les nouveaux prédicteurs que nous proposons dans ce chapitre sont performants, mais leur inconvénient est qu'il nécessite une grande quantité de données et beaucoup de temps de calcul.

4.6 Comparaison des nouveaux prédicteurs avec Q-Lasso

Comme nous l'avons dit dans la revue de la littérature sur la prédiction de délai, un an après la publication de nos travaux sur la prédiction de délai avec des algorithmes d'apprentissage machine (régression spline, réseau de neurones, krigéage stochastique), Ang et al. (2016) ont proposé un prédicteur qui utilise l'algorithme d'apprentissage du Lasso (Tibshirani, 1999). Ang et al. appellent ce prédicteur Q-Lasso. Ce prédicteur estime le temps d'attente d'un patient comme une fonction linéaire dépendant de l'état du système avec un objectif de minimiser le MSE des prédictions plus une fonction de pénalité pour éviter le sur-apprentissage. Leur définition de l'état du système à l'arrivée d'un patient est très similaire du nôtre. Ici, en plus du LES, de la longueur de la file, du nombre de serveurs, de la période de la journée, les auteurs considèrent les prédicteurs QL et la liste de

priorité des médecins (serveurs) dans la définition de l'état du système.

Avec une légère réadaptation, ce prédicteur peut bien être utilisé dans les centres d'appels multi-compétences. Toutes leurs variables d'état sont observables ou estimables dans dans les centres d'appels multi-compétences sauf le délai prédit par QL. Ce dernier n'est pas applicable dans les centres d'appels multi-compétences. Si nous éliminons le délai prédit par QL, nous obtenons une version du Q-Lasso utilisable dans les centres d'appels multi-compétences.

Dans cette section, nous comparons ce prédicteur Q-Lasso adapté aux prédicteurs ANN, SK et RS pour les modèles N courtes et longues, et pour le centre d'appels HQ, qui sont étudiés dans la section 4.3.4. Les tableaux 4.26, 4.25 et 4.27 donnent les RRASEs des divers prédicteurs pour le modèle N avec longues files, le modèle N avec courtes files et pour le modèle HQ avec 6 types d'appels et 8 groupes, basé sur des données réelles. Nous constatons que les prédicteurs ANN, RS, SK sont plus précis que Q-Lasso dans les trois exemples. La différence de précision entre Q-Lasso et RS varie entre 3% et 5% à travers les différents exemples. Cependant, si nous comparons les temps nécessaires pour l'entraînement des modèles de prédictions, nous constatons Q-Lasso a un temps d'entraînement beaucoup plus petit que ceux de ANN et SK, mais sensiblement égal à celui de RS.

Type	LES	Q-Lasso	RS	SK	ANN
1	0.871	0.635	0.603	0.590	**0.572**
2	0.850	0.625	0.597	0.591	**0.560**

Tableau 4.25 : Le RRASE pour le N. modèle avec de courtes files.

Type	LES	Q-Lasso	RS	SK	ANN
1	0.499	0.405	0.364	0.360	**0.341**
2	0.629	0.479	0.443	0.445	**0.424**

Tableau 4.26 : Les RRASEs pour le modèle N avec longues files.

Recommandations sur les prédicteurs Pour chaque type d'appel k, nous avons proposé des prédicteurs qui utilisent des algorithmes d'apprentissage machine

	Types d'appels					
	1	2	3	4	5	6
LES	24.6	35.6	20.3	41.3	26.1	94.5
Q-Lasso	12.7	16.4	14.6	19.3	22.5	65.8
RS	8.90	12.9	11.4	15.9	18.9	62.7
SK	8.50	11.6	10.1	14.0	16.5	**61.5**
ANN	**7.50**	**10.1**	**8.20**	**12.2**	**14.8**	63.7

Tableau 4.27 : Le RRASE des prédicteurs pour les 6 types d'appels du centre d'appels HQ.

et qui prennent en entré un vecteur $\boldsymbol{x}$ qui est constitué du délai LES, de la longueur de la file du type k, de la longueur des autres types servis par les mêmes agents. Dans tous exemples considérés, nous constatons que les nouveaux prédicteurs sont performants. Dans les systèmes à file d'attente unique avec agents identiques et des durées de service exponentielles, nos prédicteurs sont aussi performants que le prédicteur optimal QL. Pour les systèmes à file unique avec des agents hétérogènes et des durées de service de loi log-normale, nos prédicteurs sont largement plus performants que QL.

Les performances de nos prédicteurs peuvent êtres améliorées en ajoutant le staffing des groupes, la période d'arrivée des appels dans $\boldsymbol{x}$. Dans tous les exemples considérés, nous observons une réduction du RRASE qui varie entre 1 et 5 %. Cependant, il faut noter que l'ajout de ces informations peut doubler ou tripler le temps d'entraînement nécessaire pour trouver les bons paramètres des prédicteurs. Nous avons constaté que plus la taille du vecteur $\boldsymbol{x}$ est grande, plus le temps nécessaire pour apprendre le bon vecteur de paramètres θ augmente. Le temps d'entraînement dépend aussi de la taille des données disponibles. Si la taille des données est très grande, l'entraînement des modèles SK, ANN peut prendre plusieurs journées. Pour réduire le temps d'entraînement, nous utilisons une technique d'agrégation des données présentée à la section 4.2.3.

En général, ANN est le prédicteur le plus précis, mais il demande le plus souvent plusieurs essais pour trouver la bonne configuration du réseau (combien de couches cachées, combien neurones par couche). Nous ne connaissons pas une méthode

pour déterminer la configuration optimale du réseau. Le temps nécessaire pour trouver les bons paramètres peut souvent dépasser plusieurs heures surtout si la taille du vecteur $\boldsymbol{x}$ est grande. Si la taille des données d'entraînement est petite, ANN performe moins bien. Dans les différents exemples étudiés dans ces situations, nous avons observé qu'il est préférable d'utiliser RS

SK est la plupart du temps moins bon que ANN et en plus le temps nécessaire pour entraîner le modèle est beaucoup plus grand que celui de ANN. Il est en général plus précis que RS, mais la différence de précision varie le plus souvent entre 1 et 2%.

RS est souvent le prédicteur le moins précis, mais son temps d'entraînement est largement plus petit que ceux de ANN et SK. Il a l'avantage d'être rapide pour trouver le bon vecteur de paramètres θ. En général quelques minutes suffisent pour le trouver. Il est aussi le prédicteur le plus précis si la taille des données d'entraînement est petite.

Le choix d'un prédicteur par un gestionnaire dépend du niveau de la précision des prédictions voulue, du temps dont on dispose pour entraîner le modèle, et de la quantité de données disponibles.

CHAPITRE 5

NOUVEAUX PRÉDICTEURS DE DÉLAI BASÉS SUR L'HISTORIQUE POUR LES SYSTÈMES DE SERVICE

5.1 Introduction

Dans ce chapitre, nous présentons les travaux de notre article Thiongane et al. (2016). Nous proposons deux nouveaux prédicteurs de délais qui sont très simples à mettre en œuvre et peuvent être utilisés dans les systèmes multi-compétences. Ils sont basés sur les temps d'attente des clients précédents de la même classe. Le premier estime le délai d'un nouveau client en extrapolant l'historique des attentes des clients actuellement dans la file d'attente, plus le dernier qui est entré en service, et en prenant une moyenne pondérée. Le second prend une moyenne pondérée des délais des anciens clients de la même classe qui ont trouvé la même longueur de file d'attente quand ils sont arrivés.

5.1.1 Contexte et problème

Nous allons présenter le contexte et les problèmes qui motivent la proposition de nouveaux prédicteurs. Ils dépendent du type de système considéré.

Systèmes avec file unique

Nous rappelons que deux catégories de prédicteurs ont été développées pour les systèmes à file unique : les prédicteurs QL et les prédicteurs DH.

Les prédicteurs QL sont connus pour être optimaux pour les systèmes simples comme une file d'attente M/M/s ou une file d'attente M/M/s+M à l'état d'équilibre pour lesquels les agents sont identiques et les durées de service exponentielles. Dans les centres d'appels, des études empiriques menées sur de vraies données ont montré que les serveurs sont hétérogènes et les temps de service sont de loi log-normale plutôt qu'exponentiels (Armony, 2005, Gans et al., 2010, Mehrotra et al.,

2012, Pichitlamken et al., 2003). L'utilisation des prédicteurs QL qui ignorent ces réalités peut conduire à de très mauvaises prédictions.

Les prédicteurs DH sont performants dans les systèmes avec une faible variation dans les processus d'arrivée et une faible variation du nombre de serveur. Leurs performances se dégradent considérablement s'il y a une grande variation du processus d'arrivée ou du nombre de serveurs dans le temps. Malheureusement dans les centres d'appels actuels ces variations sont importantes.

Systèmes multi-compétences

Dans les systèmes multi-compétence, les prédicteurs QL ne sont pas applicables. Les prédicteurs DH donnent de mauvaises performances, car les variations des processus d'arrivée et du nombre de serveurs dans le temps sont les caractéristiques des centres d'appels multi-compétences actuels. D'autres types de prédicteurs, qui appliquent généralement des algorithmes d'apprentissage machine sur les données observées, sont proposés pour ces systèmes (Ang et al., 2016, Senderovich et al., 2015, Thiongane et al., 2015). Ces prédicteurs performent bien empiriquement dans les simulations, mais un inconvénient est qu'ils ont un grand nombre de paramètres qui doivent être appris à l'avance, et ils sont complexes à mettre en œuvre en pratique. Cette phase d'entraînement du modèle nécessite une grande quantité de données et temps de calcul.

5.1.2 Objectifs

Dans ce chapitre, nous nous concentrons sur les prédicteurs simples de type DH qui sont faciles à mettre en œuvre et qui ont très peu de paramètres. Nous proposons deux nouveaux prédicteurs. Le premier étend le prédicteur LES en considérant les temps d'attente vécus jusqu'ici par les clients de même type qui sont *encore dans la file d'attente*. Les temps d'attente finaux de ces clients sont encore inconnus, mais le prédicteur extrapole les temps d'attente qu'ils ont réalisé jusqu'à présent. Nous appelons ce prédicteur le *LES extrapolé* ou *extrapolated LES* (E-LES). Le second

prédicteur estime le temps d'attente du nouveau client par une moyenne mobile des temps d'attente réalisés des clients de même type qui ont trouvé la même longueur de file d'attente quand ils sont arrivés. Nous l'appelons la *moyenne conditionnelle des LES* ou *average conditional LES* (AvgC-LES). Ces nouveaux prédicteurs sont attrayants en grande partie en raison de leur simplicité. En effet, ils ont très peu de paramètres, ne nécessitent pas une phase d'optimisation, et sont faciles à mettre en œuvre en pratique. Le second a un seul paramètre : la taille de la fenêtre pour la moyenne mobile. Le premier n'en a pas dans sa forme de base, alors que certaines de ses variantes ont un paramètre qui sert à exclure certains des clients à l'arrière de la file d'attente, dont le temps d'attente réalisé jusqu'à présent ne fournit pas suffisamment d'information. Nous étudions ces prédicteurs dans le contexte des centres d'appels, mais ils peuvent également être utilisés pour d'autres systèmes de services.

Nous avons effectué des expériences par simulation pour comparer la précision des différents prédicteurs sur différents modèles de centres d'appels. Dans ces expériences, nous avons constaté que AvgC-LES était généralement plus précis que E-LES, qui était à son tour plus précis que LES. Pour une seule file d'attente, pour laquelle QL est connu pour être optimal (si nous considérons la supposition non réaliste que les serveurs sont identiques et les temps de service sont exponentiels), AvgC-LES est très proche de QL. Dans le cas d'une seule file avec des serveurs hétérogènes et des durées de service de loi log-normale, E-LES et AvgC-LES sont plus performants que QL. Pour les exemples de centre d'appels multi-compétences, AvgC-LES est un peu moins précis que RS et ANN. Cependant, il est beaucoup plus simple.

5.1.3 Organisation du chapitre

Le reste du chapitre est structuré comme suit. La section 5.2 introduit nos nouveaux prédicteurs de délai. La section 5.3 présente les résultats des expériences numériques. Une conclusion est donnée dans la section 5.4.

5.2 Les prédicteurs de délais

Dans cette section, nous présentons les nouveaux prédicteurs de délai que nous proposons dans ce chapitre. Puisque nous sommes intéressés à des prédicteurs qui sont susceptibles d'être mis en œuvre dans la pratique, nous avons proposé des prédicteurs DH qui ont très peu de paramètres. À titre de comparaison, dans nos études de simulation, nous incluons LES, Avg-LES, P-LES, QL pour les systèmes avec une seule file d'attente, et les prédicteurs qui utilisent les algorithmes d'apprentissage machines (ML) pour les instances multi-compétence à plusieurs files d'attente. Notez que même si QL et ML sont plus performants (quand ils sont applicables), ces prédicteurs ont d'autres limites, comme nous l'avons dit un peu plus tôt dans l'introduction. Les prédicteurs DH utilisent toujours les délais des clients de la même classe (même file d'attente) que celle pour laquelle nous faisons la prédiction. Ainsi, pour chaque méthode considérée, il y a un prédicteur différent pour chaque catégorie de clients j, même si nous ne l'indexons pas toujours par j explicitement. Par exemple, si nous utilisons LES dans un modèle N, nous aurons un prédicteur LES pour le type d'appel 1 et un prédicteur pour le type d'appel 2.

5.2.1 LES extrapolé (E-LES)

Ici, nous proposons un prédicteur de DH qui repose sur des informations de délai des clients qui sont *actuellement en attente dans la file*. Les délais finaux de ces clients sont encore inconnus, mais nous extrapolons les délais (partiels) écoulés pour les prédire. Ceci est la principale distinction entre E-LES et les prédicteurs DH précédents (LES, Avg-LES, WAvg-LES et P-LES), qui reposent uniquement sur les délais passés qui sont déjà complets. E-LES utilise des informations partielles, mais plus fraîches. Il fonctionne comme suit.

Supposons qu'un nouveau client entre dans une file d'attente avec C clients avant lui, numérotés de 1 à C, avec le client 1 à la tête de la file d'attente. Le client LES, qui était juste en avant du client 1, a le numéro 0. Pour tout client $c \in \{1, \ldots, C\}$, soient $Q(c)$ le nombre de clients *déjà en file d'attente* lorsque le

client c arrive, $A(c)$ le nombre de clients *actuellement* devant c et $W(c)$ le temps d'attente vécu par le client c jusqu'à maintenant. Ainsi, le client c a trouvé $Q(c)$ clients dans la file d'attente à l'arrivée, et a progressé de $Q(c) - A(c)$ positions dans la file d'attente pendant le temps écoulé $W(c)$ depuis son arrivée. Sachant que $Q(c) + 1$ clients doivent sortir du système (après avoir été servis ou avoir abandonné) avant que client c puisse commencer le service, il semble naturel de prédire le temps d'attente $E(c)$ du client c par l'extrapolation linéaire

$$E(c) = W(c)\frac{Q(c)+1}{Q(c)-A(c)}. \tag{5.1}$$

Pour $c = 0$, nous fixons $E(0)$ égal au temps d'attente réel du client LES, parce que son vrai délai est déjà connu.

Le délai prédit D du nouveau client est la moyenne des délais extrapolés des C clients dans la file d'attente et le vrai délai du client LES :

$$D = \frac{1}{C+1}\sum_{c=0}^{C} E(c). \tag{5.2}$$

La formule (5.2) basée sur (5.1) fournit une moyenne pondérée naturelle qui met plus de poids sur les délais les plus récents, donc on peut espérer qu'il capture les changements dans les systèmes dynamiques plus tôt que les autres prédicteurs de DH. Notez que parce que les C clients partagent la même file d'attente, leurs temps d'attente sont généralement corrélés. Alors les $E(c)$ ne sont pas indépendants. Une faiblesse du prédicteur (5.2) est que les clients près de la fin de la file d'attente ont vécu seulement un court temps d'attente jusqu'à présent et ont généralement une petite valeur de $Q(c) - A(c)$, d'où ils sont susceptibles de fournir moins d'informations de délai et leurs délais extrapolés $E(c)$ ont généralement du bruit. Pour réduire ce bruit, nous pouvons ajouter des poids multiplicatifs qui diminuent avec c, comme dans WAvg-LES. Ces poids peuvent dépendre de C, c, $Q(c)$ et $A(c)$.

Nous avons implémenté une version de ceci qui sélectionne un paramètre de seuil τ et inclut dans (5.2) uniquement les clients qui ont progressé d'au moins τ

positions dans la file d'attente depuis leur arrivée. Autrement dit, nous définissons $\mathcal{C} = \{c \leq C : Q(c) - A(c) \geq \tau\} \cup \{0\}$ et

$$D = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} E(c), \tag{5.3}$$

où $|\mathcal{C}|$ est la taille de l'ensemble $\mathcal{C}$. Les poids sont de $1/|\mathcal{C}|$ pour $c \in \mathcal{C}$ et 0 sinon. Dans notre implémentation, nous sélectionnons une constante fixe $\delta \in (0, 1]$ et nous prenons un seuil dynamique τ comme une proportion δ de la longueur de la file courante C, qui est $\tau = \lceil \delta Q \rceil$. Un grand δ retourne des prédictions proches de celles de LES. En particulier, si nous prenons $\delta = \infty$, alors $\mathcal{C} = \{0\}$ et nous obtenons LES. Dans nos expériences de simulation, le meilleur δ que nous avons trouvé empiriquement (à partir de quelques sélections) n'a jamais dépassé 0.4. Nous rapportons et utilisons cette meilleure valeur pour chaque exemple. Pour le premier exemple, $\delta = 0.1$, pour le second exemple $\delta = 0.2$, et pour le troisième exemple $\delta = 0.4$.

Une autre heuristique pour sélectionner les poids est de toujours inclure une proportion fixe $\beta \in [0, 1]$ des clients à partir de la tête de la file d'attente. Nous remplaçons C par $C' = \lceil \beta C \rceil$ en (5.2), qui donne le prédicteur

$$D = \frac{1}{C' + 1} \sum_{c=0}^{C'} E(c). \tag{5.4}$$

Choisir $\beta = 0$ exclut tous les clients dans la file d'attente, alors dans ce cas E-LES devient LES. Nous allons utiliser (5.3) et non (5.4) pour nos expériences de simulation.

5.2.2 Moyenne des LES conditionnelles à la longueur de la file d'attente (AvgC-LES)

Cette méthode est inspirée par le prédicteur QL pour une seule file d'attente avec des temps de service exponentiels, qui prédit le délai comme une espérance du temps d'attente conditionnelle à la longueur de la file d'attente lorsque le client

arrive. Au lieu d'utiliser une formule mathématique basée sur les temps de service exponentiels comme dans QL, le prédicteur proposé utilise les temps d'attente des clients passés de même type qui ont trouvé la même longueur de file d'attente quand ils sont arrivés.

Plus précisément, pour chaque file d'attente j, nous choisissons la taille de file d'attente maximale K_j à considérer et, pour chaque taille de file d'attente $k \in \{1, \ldots, K_j\}$, nous choisissons un entier $N_{j,k} > 0$ comme dans Avg-LES. Nous mémorisons les temps d'attente des $N_{j,k}$ derniers clients de la classe j qui ont trouvé une file d'attente de taille k à leur arrivée. Pour une nouvelle arrivée de type j qui trouve une file d'attente j de taille k, le temps d'attente est prédit par la moyenne de ses $N_{j,k}$ temps d'attente précédents. Si k est non borné ou si certaines valeurs de k sont rares, alors nous pouvons regrouper les valeurs dans un petit nombre de sous-ensembles et de maintenir une moyenne pour chaque sous-ensemble. Si moins de $N_{j,k}$ attentes ont été enregistrées jusqu'à présent, nous prenons la moyenne de ceux enregistrés. Si aucun n'a été enregistré, nous prenons LES.

Pour ce prédicteur, contrairement à Avg-LES, un grand $N_{j,k}$ est généralement beaucoup mieux que $N_{j,k} = 1$. La principale différence avec Avg-LES est qu'ici la moyenne est seulement sur les clients qui voient la même longueur de file d'attente quand ils arrivent. Nous avons observé que pour de longues simulations avec une seule file d'attente, la précision de ce prédicteur est très proche de celle de QL. Ceci peut être expliqué par le fait que AvgC-LES a collecté suffisamment de données pour calculer une bonne espérance des temps d'attente conditionnels comme QL. Dans un régime à trafic intense avec de nombreux serveurs (Whitt, 2004), AvgC-LES avec $N_{j,k} = 1$ devient le prédicteur LES conditionnel à la longueur de la file d'attente, dont les prédictions sont proches de celles de QL.

On pourrait également envisager des versions avec moyennes pondérées de AvgC-LES, qui remplacent la moyenne ordinaire des $N_{j,k}$ précédents temps d'attente pour la classe j et la taille de la file k par une moyenne pondérée comme dans WAvg-LES. En particulier, on peut utiliser des poids décroissants exponentiellement avec de petits facteurs de lissage (par exemple, 0.1 ou moins), de sorte que

chaque nouvelle observation fait une contribution relativement faible à la moyenne. Un avantage du lissage exponentiel sur le long terme est qu'on n'a pas besoin de stocker toutes les observations individuelles des temps d'attente. Dans nos expériences, le lissage exponentiel est similaire à la moyenne ordinaire, mais ne fait jamais mieux que ce dernier en termes d'erreur de prédiction, ainsi nous ne rapportons pas les résultats détaillés pour elle.

5.3 Les résultats des simulations

Dans cette section, nous présentons les résultats des expériences de simulation qui comparent la précision des anciens et nouveaux prédicteurs sur quatre modèles de files d'attente. Nous commençons avec le modèle classique M/M/s+M, pour laquelle une formule analytique est disponible pour l'espérance du délai conditionnelle l'état actuel du système. Dans cet exemple, nous supposons les s serveurs sont tous identiques. Le but est de vérifier que nos prédictions ne sont pas trop loin de ces espérances exactes dans ce cas simple. Notre deuxième exemple est un modèle M/LN/s+M plus réaliste avec des durées service de loi log-normale et des agents hétérogènes. Nous avons une distribution des temps de service de loi log-normale pour chaque agent du groupe. Le troisième exemple est un modèle N, avec deux catégories de clients et deux groupes de serveurs, dans lequel le premier groupe sert uniquement les clients de la première classe et le deuxième groupe sert les deux classes. Le quatrième est un modèle d'un centre multi-compétence basé sur des données réelles du centre d'appel d'un fournisseur de services publics au Québec, Canada. Le modèle comporte six catégories de clients (appelés types), huit groupes d'agents, et est non-stationnaire.

Pour les prédicteurs qui nécessitent des paramètres, nous avons exploré quelques choix et sélectionné ceux qui ont donné les meilleurs résultats. Le meilleur N_j pour Avg-LES est généralement de petite taille (moins de 10 et souvent égal à 1) et les meilleurs $N_{j,k}$ pour AvgC-LES sont généralement de grande taille (100 ou plus). En accord avec cela, nous avons trouvé dans nos expériences que, pour le lissage

exponentiel, le meilleur facteur de lissage α_j est généralement plus grand que 0.9 pour ESAvg-LES et inférieur à 0.1 pour la version exponentielle pondérée de AvgC-LES. Étant donné que les résultats étaient également très semblables à ceux de la moyenne ordinaire, nous ne les rapportons pas dans les tableaux.

5.3.1 Une file d'attente avec un unique type M/M/s+M

Nous considérons un modèle de file d'attente unique M/M/s+M avec un taux d'arrivée variable dans le temps. La journée est divisée en 20 périodes d'une heure. Le processus d'arrivée est Poisson avec un taux λ_p constant à la période p, pour $p = 1, \ldots, 20$. Nous prenons $\lambda_p = 25$ pour p impair et $\lambda_p = 20$ pour p pair. Les temps de service sont exponentiels de moyenne 1 et les temps de la patience sont exponentiels avec une moyenne de 2. Il y a $s = 20$ serveurs pour toute la journée. Nous simulons 100 jours indépendants du modèle pour estimer la précision des prédicteurs. Nous constatons que la longueur de la file d'attente moyenne au cours de la journée est de 7.7 clients, la probabilité de délai est 91.9%, la probabilité d'abandon est de 15.8 %, et le temps d'attente moyen est de 1188 secondes.

Pour ce modèle, le prédicteur QL (3.10) donne l'espérance conditionnelle exacte et minimise le MSE, il est donc optimal pour notre critère, sous l'hypothèse des temps de service et temps de patience exponentiels avec des moyennes connues et constants (c-à-d si μ^{-1}, ν^{-1} et $s = 20$ sont connus et ne varient pas avec le temps). Nous comparons les performances des autres prédicteurs avec QL pour voir à quel point ils sont proches d'être optimaux.

Le tableau 5.1 rapporte les RRASEs pour les divers prédicteurs. Nous avons utilisé $N_j = 2$ pour Avg-LES, $N_{j,k} = 100$ pour AvgC-LES, et $\delta = 0.1$ pour E-LES. QL gagne, ce qui est sans surprise, suivi de très près par AvgC-LES. Les autres méthodes donnent des RRASEs beaucoup plus grands, et le meilleur d'entre eux est notre nouveau prédicteur E-LES. Avg-LES avec $N_j \geq 2$, souvent utilisés dans la pratique, fait pire que LES, qui correspond à $N_j = 1$. Ibrahim et al. (2016a) ont trouvé un comportement similaire. P-LES se révèle être le plus mauvais prédicteur. La figure 5.1 donne un histogramme des erreurs de prédicteurs pour les prédicteurs

LES, E-LES, AvgC-LES et QL. Nous observons que le prédicteur LES a une distribution d'erreur beaucoup plus large (les grandes erreurs sont plus fréquentes), alors que AvgC-LES et QL ont des distributions d'erreur très similaires.

Tableau 5.1 : RRASEs for the M/M/20+M example.

	LES	Avg-LES	P-LES	E-LES	AvgC-LES	QL
RRASE	46.9	49.4	59.2	43.6	32.9	**32.1**

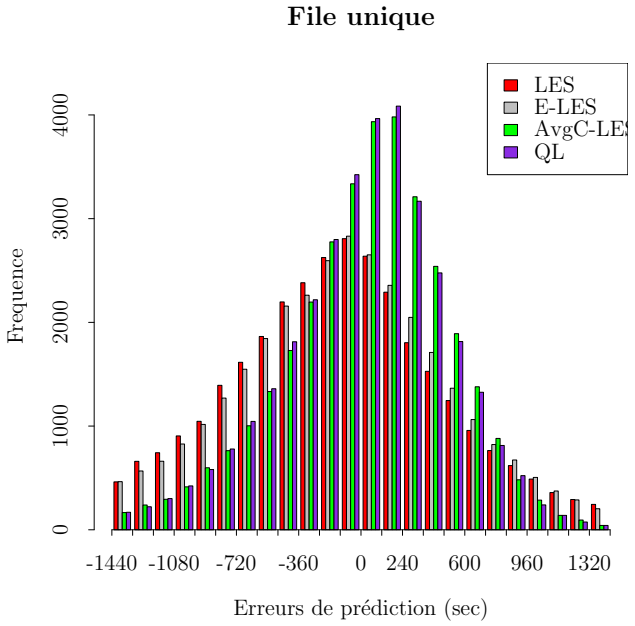

Figure 5.1 : Modèle M/M/s+M : distribution de l'erreur de prédiction.

Les figures 5.3.1 – 5.3.1 affichent les délais réels et les délais prédits par QL, LES, AvgC-LES, et P-LES, en fonction de l'heure d'arrivée, pour quatre journées distinctes. Il donne une idée de comment se comportent les erreurs de prédiction. Elle montre que QL est la plupart du temps meilleur que AvgC-LES, mais parfois leurs prédictions sont très proches. Elle montre aussi que AvgC-LES est souvent meilleur que LES et P-LES. P-LES montre une plus grande volatilité, car il réagit plus rapidement au bruit stochastique, mais cela peut parfois conduire à de grandes erreurs de prédiction. Bien sûr, ce comportement diffère selon les différents jours.

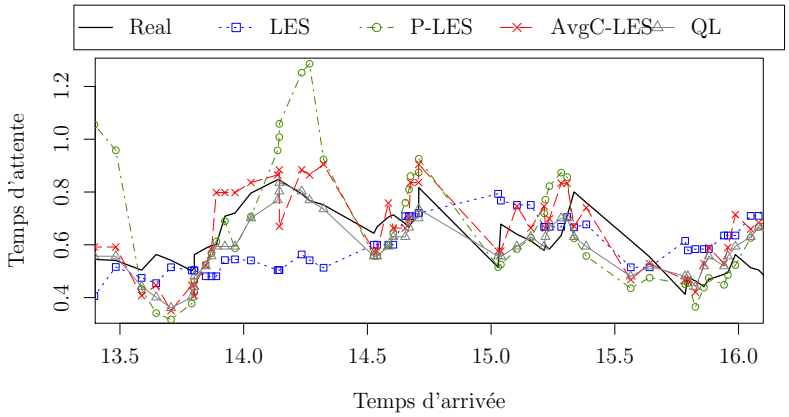

Figure 5.2 : Jour 1

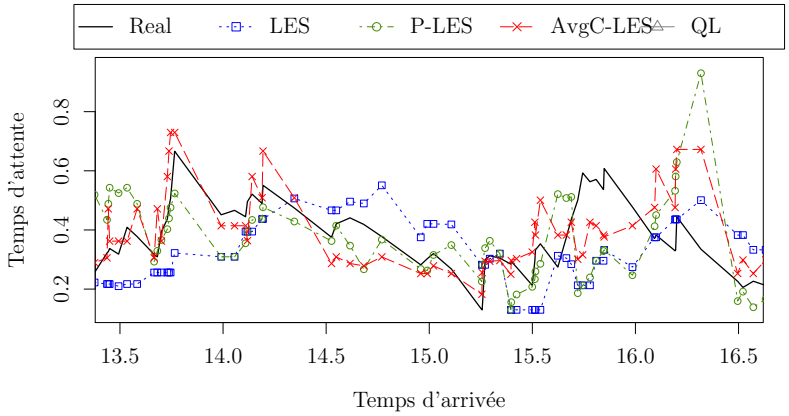

Figure 5.3 : Jour 2

5.3.2 Une file d'attente unique de type M/LN/s+M

Nous allons maintenant comparer les performances des différents prédicteurs dans un système de file d'attente avec des serveurs hétérogènes et des temps de service de loi log-normale. Nous considérons le même exemple étudié à la section 4.3.3 du chapitre 4.

Le tableau 5.2 rapporte les RRASE×100 des différents prédicteurs. Contrairement à l'exemple précédent, nous constatons ici que AvgC-LES est largement plus précis que QL et suit de près les prédicteurs ANN et RS qui nécessitent une phase

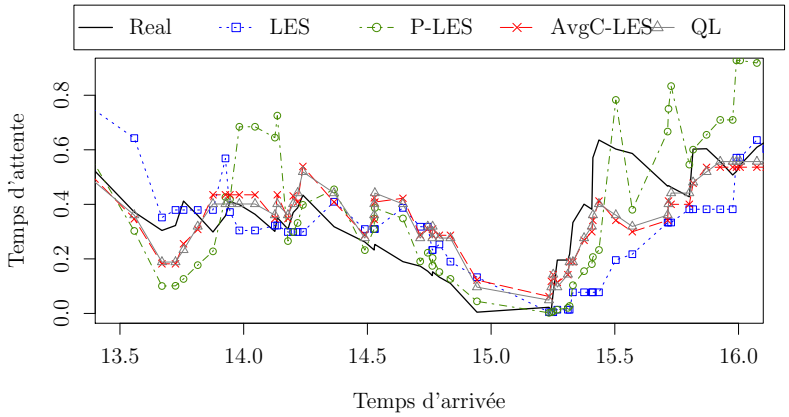

Figure 5.4 : Jour 3

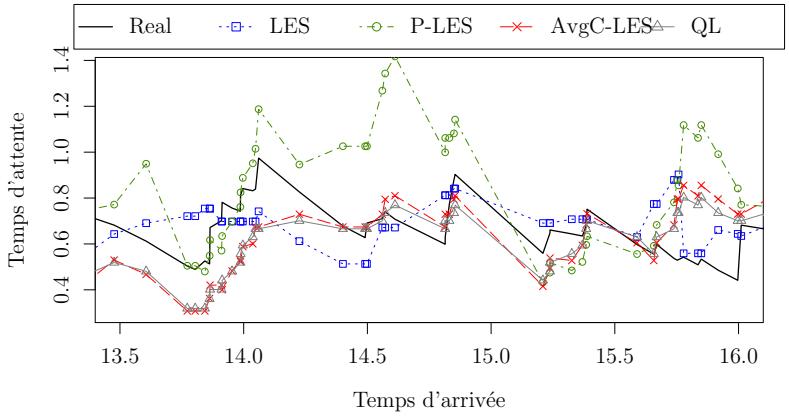

Figure 5.5 : Jour 4

d'entraînement et des données. Il est suivi par E-LES, LES, Avg-LES, QL et P-LES respectivement. Comme dans le précédent exemple, la différence de précision

	LES	Avg-LES	P-LES	E-LES	AvgC-LES	QL	RS	ANN
RRASE	30.6	31.2	35.3	27.5	17.8	31.8	16.5	**14.3**

Tableau 5.2 : RRASE des prédicteurs pour le modèle M/M/10+M avec des taux d'arrivée variable dans le temps.

entre AvgC-LES et E-LES est grande. Ici elle est d'environ 10%. La différence de précision entre E-LES et LES reste sensiblement la même que dans l'exemple avec

des temps de service exponentiels (environ 3%). Entre QL et LES, cette différence est d'environ 1% . Les mauvaises performances de QL s'expliquent par le fait qu'il n'est pas adapté pour les modèles avec des agents hétérogènes et des temps de service de loi log-normale.

Une remarque importante est la suivante : l'utilisation d'agents hétérogènes et des durées de service de loi log-normale à la place des agents homogènes et des temps de service exponentiels dans la modélisation des centres d'appels n'affecte pas la précision des prédicteurs AvgC-LES et E-LES, mais affecte considérablement la précision du prédicteur QL.

5.3.3 Modèle N de centre d'appels

Nous considérons différents exemples de modèle N. Plus précisément, nous utilisons 4 exemples de centres d'appels d'un modèle N. Dans les deux premiers exemples, nous utilisons des modèles avec file d'attente courte et dans les deux derniers exemples, nous utilisons des modèles avec de longues files d'attente. Dans chaque type de système (courtes files ou longues files), nous étudierons deux cas. Le premier avec des agents homogènes et des durées de service exponentielles, et le second avec des agents hétérogènes et des durées de service de loi log-normale. Le premier cas est le plus souvent utilisé dans la modélisation des centres d'appels, mais le second s'ajuste mieux aux données réelles des centres et il est donc le plus réaliste.

Outre les prédicteurs de DH, nous avons aussi essayé les prédicteurs RS et ANN de Thiongane et al. (2015), mentionné dans l'introduction. Les prédicteurs RS et ANN sont les plus performants, mais ils ont besoin d'une phase d'apprentissage qui est très coûteuse et ont de nombreux paramètres. Nous les utilisons comme benchmark pour la comparaison.

5.3.3.1 Un modèle N avec courte file d'attente

Ici nous utilisons deux exemples de modèle N avec files d'attente courtes. Le premier avec des agents identiques et des temps de service exponentiels. Le second avec des agents hétérogènes et des durées de service de loi log-normale.

Agents homogènes et durées de service exponentielles

Nous considérons le même exemple avec des agents identiques et les temps de service exponentiels qui est étudié à la section 4.3.4.1.

Le tableau 5.3 rapporte les RRASEs pour les deux types d'appels. Comme prévu, les résultats de RS et ANN sont meilleurs que ceux des prédicteurs DH. AvgC-LES est de loin le prédicteur DH le plus précis. Il est suivi par E-LES, LES, Avg-LES, et P-LES, respectivement. Lorsque nous comparons les RRASEs de AvgC-LES avec ceux de RS et ANN, nous constatons que la différence de précision n'est pas trop grande aussi bien pour le type 1, et le type 2. Le RRASE de AvgC-LES est environ 4% plus élevé pour le type 1 et environ 2% plus élevé pour le type 2. Par contre si nous comparons le RRASE de AvgC-LES avec celle de E-LES, nous observons que la différence est grande. Le RRASE de ce dernier est d'environ 21% plus élevé que celui de AvgC-LES dans tous les cas. Nous observons que les résultats de E-LES et de LES sont proches. Ceci était prévisible, car nous avons des files d'attente courtes qui contiennent en moyenne 1 ou 2 clients, et nous savons que E-LES, pour diminuer le bruit dans les prédictions, ne considère pas les délais extrapolés des clients qui sont à la fin de la file et qui ont avancé que très peu de positions depuis leur entrée dans la file. Ainsi, nous aurons la plupart du temps une prédiction de E-LES égale à celle de LES où bien une prédiction très similaire à celle de LES.

Type	LES	Avg-LES	P-LES	E-LES	AvgC-LES	RS	ANN
1	87.1	88.4	123	86.2	64.8	60.3	**59.0**
2	85.0	86.4	117	84.6	61.5	59.7	**57.1**

Tableau 5.3 : RRASE pour chaque type d'appel, pour le modèle N avec courtes files.

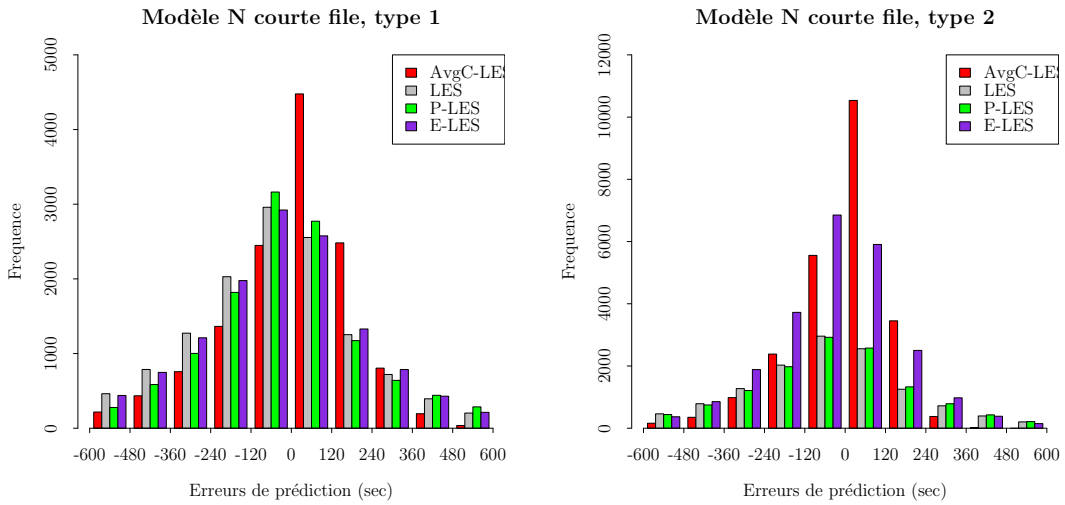

Figure 5.6 : Modèle N avec courtes files : Distribution des erreurs de prédictions (délai estimé moins délai réel) pour le type 1 et le type 2.

Agents hétérogènes et durées de service de loi log-normale

Nous considérons maintenant que les agents sont hétérogènes et les durées de service de loi log-normale. Chaque agent a sa propre distribution pour les temps de service. Nous utilisons le même exemple que précédemment à la seule différence que maintenant les durées de service sont de loi log-normale. Maintenant, nous considérons que les agents sont hétérogènes, mais la moyenne des durées de service de l'ensemble des agents est toujours la même que dans l'exemple précédent.

Pour le type 1, la moyenne des durées de service, pour l'ensemble des agents qui peuvent le servir, est toujours de 20 minutes comme dans le cas avec des agents homogènes. Chaque agent i a une moyenne m_i qui est comprise entre 18 et 22 minutes. Nous supposons que la variance des durées de service $v_i = 400$. Elle est la même pour tous les agents (nous prenons la même variance que dans le cas des durées de service exponentielles). Pour le type 2, la moyenne des durées de service pour tous les agents est 10 minutes, chaque agent a sa propre moyenne m_i comprise entre 8 et 12 minutes et une variance $v_i = 100$.

Nous simulons une journée du centre d'appel avec 100 réplications. Le partage

des compétences est très présent : 82% des appels servis de type 1 sont servis par le groupe 1 et les 18% restants sont servis par le groupe 2. Le tableau 5.4 donne les mesures de performance agrégées sur toutes les périodes pour les deux types d'appels. Nous constatons que dans cet exemple, les mesures de performances agrégées sont beaucoup plus petites que dans l'exemple avec des agents homogènes. Une probabilité de délai plus petit qui entraîne naturellement moins d'abandons et des longueurs de file d'attente moyenne plus courtes.

Performance mesures	Type 1	Type 2
PD (%)	35.9	54.4
PA (%)	6.24	8.39
AQS	0.68	1.06
AWT (Sec.)	90	101
AWT' (Sec.)	263	185

Tableau 5.4 : Mesures de performance moyenne du modèle N courtes files avec agents hétérogènes et temps de service de loi log-normale.

Le tableau 5.5 donne les RRASEs pour les deux types d'appels. Sans surprise, RS et ANN donnent les meilleurs résultats. Parmi les prédicteurs DH, AvgC-LES est le prédicteur le plus précis, mais cette fois ci sa précision a considérablement diminué alors que celle de LES est restée sensiblement la même. Pour le type 1, le RRASE de AvgC-LES a augmenté de 12% alors que celui de LES a augmenté de 2%, et pour le type 2, ces augmentations sont de 8% pour AvgC-LES et de 1% pour LES. Cette perte de précision de AvgC-LES s'explique par le fait que la probabilité de délai est faible (il y a eu peu d'attente) dans le système, et de ce fait il y aura un trop petit historique de délais pour calculer de bonnes espérances conditionnelles des temps d'attente.

Nous constatons que LES et E-LES donnent sensiblement les mêmes résultats. Ceci était prévisible vu que les longueurs moyennes des files d'attente sont de 0.68 et 1 clients alors dans ces conditions E-LES devient LES. Avg-LES suit de très près LES et comme dans les autres exemples, P-LES est toujours le prédicteur le plus mauvais.

Type	LES	Avg-LES	P-LES	E-LES	AvgC-LES	RS	ANN
1	89.6	90.3	135	89.9	78.8	61.5	**59.1**
2	86.7	87.2	115	87.1	69.5	60.2	**58.4**

Tableau 5.5 : RRASE pour chaque type d'appel, pour le modèle N avec courtes files et des agents hétérogènes.

5.3.3.2 Un modèle N avec de longues files d'attente

Nous considérons deux modèles N avec des files d'attente longues. Comme dans le cas avec courtes files, nous allons étudier deux types d'exemples. Le premier exemple avec des agents identiques et des durées de service exponentielles et le second avec des agents hétérogènes et des temps de service de loi log-normale.

Agents homogènes et durées de service exponentielles

Nous considérons en premier que les agents sont identiques et les durées de service sont exponentielles. Nous avons utilisé l'exemple étudié à la section 4.3.4.2 du chapitre 4.

Le tableau 5.6 rapporte les RRASE×100 pour les deux types d'appels, pour divers prédicteurs. Nous avons pris $N_j = 7$ pour Avg-LES, $N_{j,k} = 100$ pour AvgC-LES, et $\delta = 0.2$ pour E-LES. Les prédicteurs RS et ANN fournissent les meilleurs résultats. Comme dans l'exemple avec courtes files, RS et ANN sont encore suivis de très près par AvgC-LES qui est le meilleure parmi les prédicteurs DH. P-LES est le plus mauvais. LES, Avg-LES, et E-LES ont des performances comparables. La figure 5.7 affiche la distribution des erreurs de prédictions. Elle confirme que AvgC-LES est le meilleur prédicteur parmi les prédicteurs DH, et montre que la différence de performance entre ce dernier et ANN n'est pas grande.

Type	LES	Avg-LES	P-LES	E-LES	AvgC-LES	RS	ANN
1	49.9	52.1	70.2	46.7	37.3	36.4	**35.1**
2	62.9	67.1	94.6	61.0	47.3	44.3	**42.3**

Tableau 5.6 : RRASE pour chaque type d'appel, pour l'exemple du modèle N.

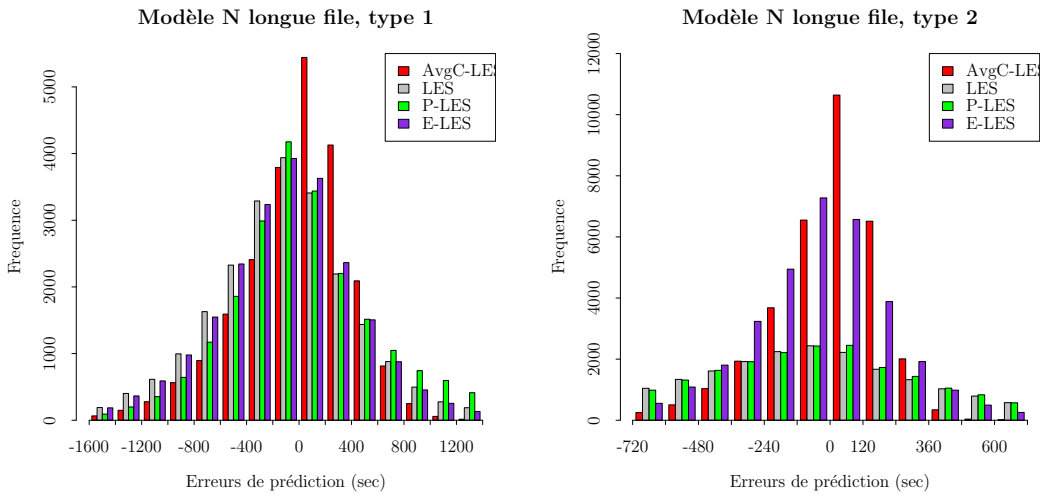

Figure 5.7 : Modèle N avec longues files : Distribution des erreurs de prédiction (délai estimé moins réel) pour les types 1 et 2.

Agents hétérogènes et durées service de loi log-normale

Nous considérons maintenant que les agents sont hétérogènes et les durées de service sont de loi log-normale. Nous considérons le même exemple que dans le cas précédent à la seule différence qu'ici les durées de service sont de loi log-normale. La moyenne des durées de service de l'ensemble des agents est toujours la même que dans l'exemple précédent. Chaque agent i a sa propre distribution pour les temps de service de paramètres κ_i et σ_i.

Pour le type 1, le taux de service moyen sur l'ensemble des agents est 21 minutes, cependant agent a sa propre distribution de moyenne m_i comprise entre 19 et 23 et une variance $v_i = 441$. Pour le type 2, les agents ont une moyenne des durées de service de 11 minutes et chaque agent i a une distribution de moyenne m_i comprise entre 9 et 13 minutes et de variance $v_i = 121$.

Nous avons simulé 100 journées indépendantes du modèle. Nous avons trouvé que le partage des compétences est très présent, car 85% des appels de type 1 sont servis par le groupe1 et les autres 15% par le groupe 2. Le tableau 5.7 montre quelques mesures de performances pour les deux types d'appels.

Le tableau 5.8 donne les RRASE×100 des prédicteurs pour les deux types d'ap-

Performance mesures	Type 1	Type 2
PD (%)	80.0	84.8
PA (%)	22	14.5
AQS	6.3	3.4
AWT(Sec.)	601	256
Cond. AWT(Sec.)	821	313

Tableau 5.7 : Mesures de performance moyennes pour l'exemple du modèle N.

pels. Comme prévu, RS et ANN donnent les meilleurs résultats. AvgC-LES, qui est le prédicteur DH le plus précis, les suit encore de très prés. La différence de précision est environ de 2% pour les deux types d'appels. E-LES arrive en quatrième position, loin dernière AvgC-LES. LES, Avg-LES ont des performances comparables. Comme toujours, P-LES donne les plus mauvais résultats.

Type	LES	Avg-LES	P-LES	E-LES	AvgC-LES	RS	ANN
1	52.8	53.8	69.9	51.3	39.3	37.3	**35.9**
2	68.3	69.4	92.4	67.5	48.9	45.5	**43.7**

Tableau 5.8 : RRASE pour chaque type d'appel, pour l'exemple du modèle N.

5.3.4 Un grand centre d'appel basé sur des données réelles

Nous considérons le même exemple étudié à la section 4.3.5 qui est basé sur les données d'un véritable centre d'appels.

Le tableau 5.9 rapporte les RRASE×100 pour les six types d'appels. Nous avons utilisé $N_j = 10$ pour Avg-LES, $N_{j,k} = 200$ pour AvgC-LES, et $\delta = 0.4$ pour E-LES. Comme prévu, RS le plus coûteux donne les meilleures prédictions pour tous les types d'appels, AvgC-LES est le meilleur (de loin) parmi les méthodes de DH, et P-LES est le moins performant. La différence de précision entre RS et AvgC-LES est plus grande ici que dans les exemples précédents, sauf pour le type d'appel 6. L'explication est que ce type d'appel a une priorité élevée pour tous les groupes qui peuvent le servir et son taux d'arrivée ne varie pas beaucoup avec le temps. Pour les autres types d'appels, nous avons une plus grande variation des taux d'arrivée et du staffing, et cela affecte la précision des prédicteurs de DH. Ibrahim et Whitt

(2009b) ont également observé que les prédicteurs DH perdent leur précision lorsque le staffing et le taux d'arrivée varient considérablement. Cependant, nous trouvons ici que AvgC-LES perd sa précision moins rapidement par rapport aux autres prédicteurs de DH.

Tableau 5.9 : RRASEs for the 6 call types of the larger example.

Type	LES	Avg-LES	P-LES	E-LES	AvgC-LES	RS
1	24.6	25.4	45.1	23.2	13.0	**8.9**
2	35.6	34.8	95.6	34.2	22.7	**12.9**
3	20.3	21.6	28.4	20.4	16.9	**11.4**
4	41.3	55.7	67.1	39.1	22.4	**15.9**
5	26.9	28.7	31.0	25.2	22.9	**18.9**
6	94.5	96.1	130	93.2	65.8	**62.7**

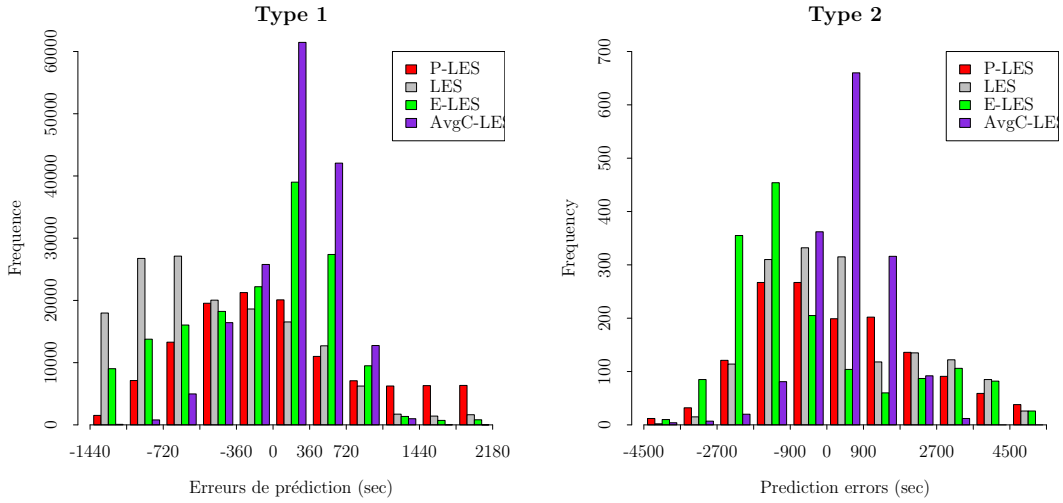

Figure 5.8 : Modèle HQ : Distribution des erreurs de prédiction pour les types 1 et 2.

Commentaires sur la robustesse de E-LES et AvgC-LES

Dans les différents exemples étudiés (file simple, modèles N courtes files, modèles N longues files et modèle HQ basé sur des données réelles), nous constatons que la présence d'agents hétérogènes, et des durées de service de loi log-normale à la place

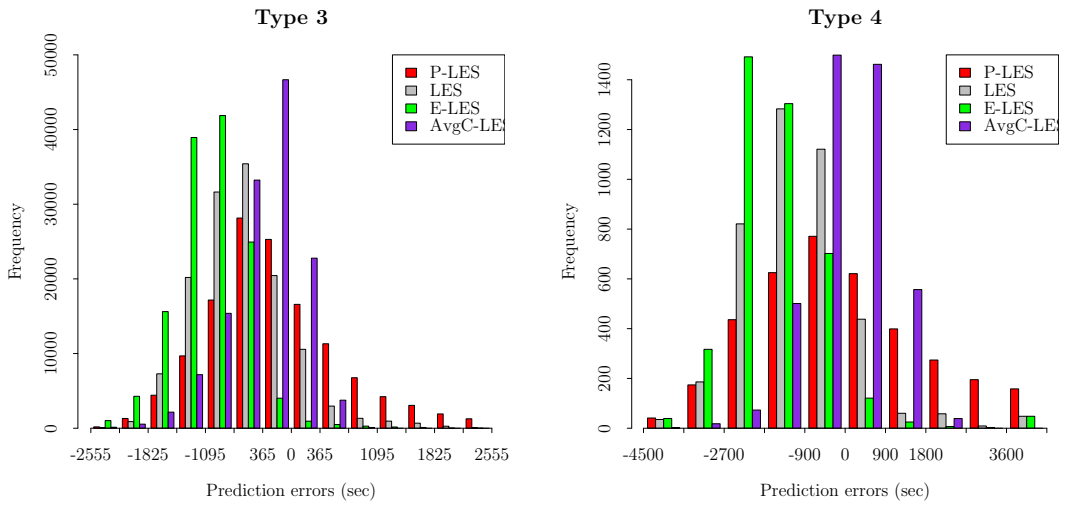

Figure 5.9 : Modèle HQ : Distribution des erreurs de prédiction pour les types 3 et 4.

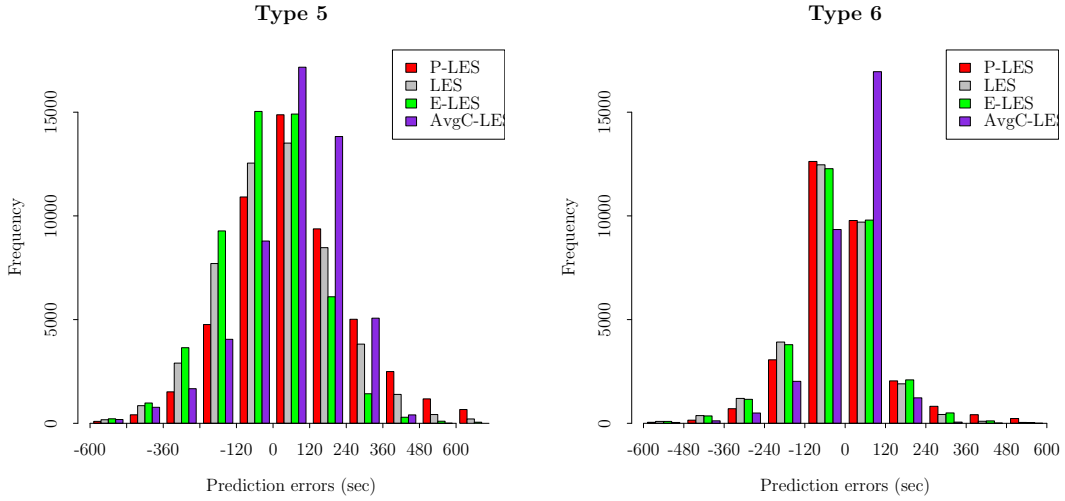

Figure 5.10 : Modèle HQ : Distribution des erreurs de prédiction pour les types 5 et 6.

d'agents homogènes et des temps de service exponentiels dans la modélisation des centres d'appels n'affecte pas la précision des prédicteurs AvgC-LES et E-LES. Ce résultat est encourageant, car dans les systèmes de la vie réelle, les temps de service sont de loi log-normale plutôt qu'exponentiels, et les agents sont hétérogènes et non

homogènes.

Pour les systèmes avec files d'attente courtes, les performances de E-LES sont sensiblement les mêmes que celles de LES car pour diminuer le bruit dans les prédictions E-LES ne considère pas les délais extrapolés des clients qui n'ont avancé que de quelques positions dans la file depuis leur arrivée. Ceci entraîne que E-LES va utiliser le plus souvent le délai LES (ou une moyenne qui est sensiblement égale au délai du client LES) pour prédire les temps d'attente des clients.

Dans les systèmes où que la probabilité de délai est faible (très peu d'attente dans le système), AvgC-LES perd de la précision (mais reste toujours meilleur que tous les autres prédicteurs DH) si les informations de délais passés qu'utilisent les prédicteurs ne sont pas assez nombreuses pour estimer de bonnes espérances conditionnelles.

5.4 Conclusion

Nous étendons la famille de prédicteurs de délais qui utilisent l'historique des systèmes de services en introduisant deux nouveaux prédicteurs de délai, basés sur des heuristiques simples. La première idée est d'exploiter l'information de délais plus récente, mais incomplète des clients toujours en attente dans la file d'attente (E-LES). Leurs temps d'attente finaux sont estimés à l'aide d'une simple extrapolation de leur progression dans la file d'attente. L'autre idée propose une version empirique de la formule QL dans le contexte des systèmes multi-compétences, en utilisant des données historiques. Pour chaque taille de file d'attente, une espérance conditionnelle des temps d'attente est estimée à partir des délais passés de clients qui ont trouvé la même longueur de file d'attente devant eux quand ils sont arrivés (AvgC-LES). Dans un système de file d'attente unique (agents homogènes et temps de service exponentiels), nos nouveaux prédicteurs sont meilleurs que les autres prédicteurs simples que nous connaissons et en plus, nous observons que AvgC-LES est très proche du prédicteur QL optimal. Dans le cas de ces systèmes avec des agents hétérogènes et des durées service de loi log-normale, E-LES et AvgC-LES en plus

d'être meilleurs que les autres prédicteurs DH, sont largement plus précis que QL. Pour les systèmes multi-compétence plus réalistes, qui ont généralement des taux d'arrivée et des staffing variables dans le temps, nos prédicteurs performent également mieux que les autres prédicteurs DH. Bien qu'ils ne battent pas les méthodes de l'apprentissage machine, leurs avantages sont qu'ils sont plus simples à mettre en œuvre, ont peu de paramètres, et ne nécessitent aucune phase d'entraînement. Ils représentent des alternatives simples et intéressantes aux prédicteurs plus complexes.

CHAPITRE 6

PRÉDICTEURS QL POUR LES CENTRES D'APPELS MULTI-COMPÉTENCES ET PRÉDICTION DE LA DISTRIBUTION CONDITIONNELLE DU DÉLAI D'ATTENTE

6.1 Introduction

Dans ce chapitre, nous proposons de nouvelles idées pour adapter les prédicteurs QL dans les centres d'appels multi-compétences. Les prédicteurs QL, qui sont développés pour les systèmes de file d'attente avec une seule file, ne sont pas directement applicables aux centres d'appels multi-compétences. Pour pouvoir les utiliser dans les systèmes multi-compétences, nous procédons comme suit. Nous proposons une représentation alternative du centre d'appels multi-compétences ayant K types d'appels en K centres d'appels indépendants. Nous définissons un centre d'appels pour chaque type d'appel et nous estimons les paramètres de chacun. Par la suite, nous utilisons les prédicteurs QL pour prédire le temps d'attente des clients qui entre aux files d'attente. Dans la deuxième partie de ce chapitre, nous parlerons de la perspective de prédire la distribution du temps d'attente du client plutôt que son espérance conditionnelle.

6.2 Nouveaux prédicteurs QL pour les centres d'appels multi compétences

Pour les centres d'appels multi-compétences étudiés dans ce chapitre, nous considérons que pour chaque type d'appel j, les arrivées se font selon un processus de Poisson de taux λ_j, et les temps de services sont des exponentiels de taux μ_j. Pour simplifier dans un premier temps, nous supposons que le taux de service dépend seulement du type d'appel et non de l'agent qui traite l'appel. Dans le cas des systèmes avec abandon, nous considérons que les temps de patience sont des variables aléatoires exponentielles de taux ν_j.

Nous faisons l'hypothèse que le centre d'appels multi-compétences peut être modélisé par un système alternatif constitué de K modèles de files d'attente indépendantes où K est le nombre de type d'appels du centre d'appels. Pour chaque type d'appel $j \in \{1, K\}$, nous avons un groupe d'agents $\tilde{G}_j$ qui traite les appels. Le nombre d'agents du groupe $\tilde{G}_j$ à la période p de la journée est égal à $\tilde{s}_{j,p}$. Nous savons que le nombre moyen d'agents qui traite un type d'appel j peut beaucoup varier dans la journée, mais nous supposons que cette variation n'est pas grande pour une période p donnée. Nous allons utiliser des heuristiques pour estimer sa valeur pour chacune des périodes.

Commençons par un exemple simple : un modèle N où nous avons 2 types d'appels et 2 groupes d'agents. Le groupe 1 (G1) peut servir uniquement les appels de type 1 et le groupe 2 (G2) peut servir les deux types d'appels. Nous représentons le modèle N par un modèle alternatif constitué de deux systèmes de files d'attente indépendants, voir figure 6.1. La première file d'attente sera constituée des appels de type 1. Nous considérons que ces derniers sont traités par les agents d'un groupe nommé $\tilde{G}_1$. Le nombre d'agents de ce groupe est $\tilde{s}_{1,p}$ à la période p. La seconde file d'attente sera constituée des appels de type 2. Les appels type 2 sont traités par les agents d'un groupe $\tilde{G}_2$ et le nombre d'agents du groupe est $\tilde{s}_{2,p}$ durant la période p.

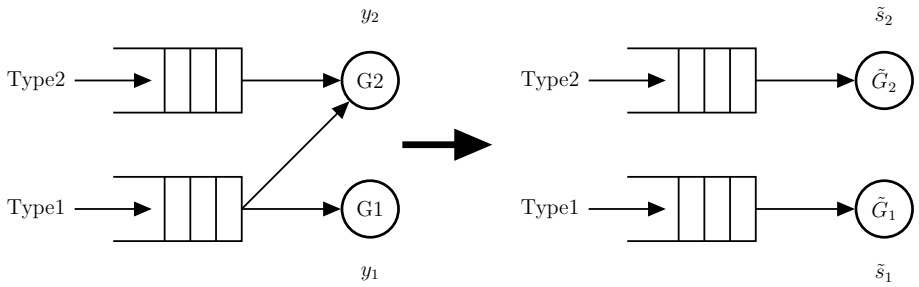

Figure 6.1 : Système alternatif du modèle N.

Pour chaque type d'appel j (chaque file d'attente), nous proposons d'utiliser un prédicteur QL_{ad} qui est une version réadaptée du QL pour un système avec une seule de file d'attente dans laquelle le nombre de serveurs s est remplacé par $\tilde{s}_{j,p}$.

Dans le cas d'un système sans abandon, le temps d'attente prédit E d'un client ayant trouvé k autres clients en attente dans la file d'attente j à la période p est prédit par

$$E = (k+1)/(\mu_j \times \tilde{s}_{j,p}) \tag{6.1}$$

Dans le cas d'un système avec abandon, le temps d'attente prédit est E par

$$E = \sum_{i=0}^{k} (\tilde{s}_{j,p}\mu + i\nu)^{-1}. \tag{6.2}$$

Dans les deux cas, $\tilde{s}_{j,p}$ (un nombre réel) peut-être toujours interprété comme étant le nombre d'agents du groupe $\tilde{G}_j$ durant la période p. Pour simplifier, nous utiliserons dans la suite $\tilde{s}$ à la place de $\tilde{s}_{j,p}$.

Pour le prédicteur QL standard (une seule file d'attente et un seul groupe d'agents) $\tilde{s}_{j,p} = s$ où s est égal au nombre total d'agents du groupe, qui est une constante connue. Par contre pour les centres d'appels multi-compétences transformés à plusieurs systèmes de files d'attente indépendantes, $\tilde{s}_{j,p}$ est inconnu et nous ne connaissons pas non plus une formule mathématique qui le détermine pour les divers groupes. Nous n'avons pas non plus proposé une formule mathématique pour déterminer $\tilde{s}_{j,p}$ à partir des paramètres du modèle, mais nous allons plutôt proposer des heuristiques pour une approximation de sa valeur. Nous allons proposer quatre méthodes pour estimer sa valeur à l'aide de données obtenues à partir d'une simulation du modèle. La première méthode utilise la moyenne empirique des temps d'attente conditionnelle à la longueur de la file d'attente k pour déterminer un $\tilde{s}_{j,p}$ pour chaque k observé durant la simulation du modèle. Ainsi, nous aurons un $\tilde{s}^k_{j,p}$ pour chaque k. La deuxième méthode utilise la moyenne pondérée de l'ensemble des $\tilde{s}^k_{j,p}$ de la première méthode pour estimer un unique $\tilde{s}_{j,p}$ pour chaque type d'appel j. La troisième méthode utilise les taux d'appels servis des groupes d'agents pour chacun des types d'appels, et le nombre d'agents des groupes pour déterminer en moyenne la proportion du nombre d'agents de chaque groupe qui sera affecté au traitement d'un type d'appel donné pour déterminer un $\tilde{s}_{j,p}$ unique pour chaque

type d'appel. La quatrième et dernière méthode utilise les proportions des temps de service des groupes à la place des taux d'appel servis pour estimer un unique $\tilde{s}_{j,p}$ pour chaque type d'appel.

6.2.1 Méthodes d'estimation du nombre d'agents des groupes

6.2.1.1 QL 1 : utilisation de la moyenne du temps d'attente conditionnelle à la longueur k de la file

Pour chaque appel de type j, nous proposons d'utiliser un $\tilde{s}^k_{j,p}$ qui est conditionnel au nombre de clients k en attente à son arrivée. Pour déterminer $\tilde{s}^k_{j,p}$ pour chaque $k \in \mathcal{N} = \{0, 1, \ldots, N\}$ où N est la longueur maximale de la file d'attente pour le type d'appel j, nous procédons comme suit.

Soit $\mathcal{C}_k = \{1, \ldots, C_k\}$ l'ensemble des clients qui ont observé une longueur de file d'attente k à leur arrivée au centre. Dans cet ensemble, nous considérons seulement les clients servis qui ont eu un temps d'attente strictement positif. Pour chaque client $c \in \mathcal{C}_k$, nous observons son temps d'attente réel $y(c)$ au début de son service. Soit W_k une variable aléatoire dénotant le temps d'attente des clients servis de type j et qui ont trouvé k clients en attente dans la file. Pour chaque k, l'espérance empirique de W_k notée $\hat{w}_k$ est estimée comme suit :

$$\hat{w}_k = \frac{1}{C_k} \sum_{c=1}^{C_k} y(c). \tag{6.3}$$

Pour chaque paire $(k, \hat{w}_k)$, nous utilisons la formule de prédiction donnée par l'équation (6.1) ou par l'équation (6.2) (le choix de la formule à utiliser dépend du modèle) pour déterminer $\tilde{s}^k_{j,p}$ conditionnel à k. Nous obtenons pour chaque type d'appel j à la période p un ensemble fini $\mathcal{D} = \{(k, \tilde{s}^k_{j,p}), \quad k = 0, ..., N\}$.

L'espérance empirique de W_k n'est rien d'autre que la valeur retournée par le prédicteur AvgC-LES (présenté au chapitre 5) pour k à la fin de la journée. AvgC-LES prend du temps avant de collecter assez d'informations pour estimer la bonne espérance conditionnelle à k alors que QL1 utilise $\tilde{s}^k_{j,p}$ qui permet d'estimer de bonne

prédiction tout au long de la journée. Mais l'avantage de AvgC-LES est qu'il n'a pas besoin de données d'historique pour être utilisé dans un centre d'appels alors QL1 requière des données pour estimer les $\tilde{s}^k_{j,p}$ avant d'être utilisé.

6.2.1.2 QL 2 : utilisation des moyennes pondérées

Avec le prédicteur QL1, pour chaque type d'appel, nous enregistrons N valeurs distinctes pour $\tilde{s}^k_{j,p}$. La méthode QL2, propose d'utiliser un unique $\tilde{s}_{j,p}$ pour chaque type d'appel j (indépendant de k) que l'on estime par la moyenne pondérée des $\tilde{s}^k_{j,p}$ de l'ensemble $\mathcal{D}$. Sa valeur est déterminée par :

$$\tilde{s}_{j,p} = \frac{1}{\sum_{k=0}^{N} C_k} \sum_{k=0}^{N} C_k \tilde{s}^k_{j,p} \tag{6.4}$$

6.2.1.3 QL 3 : utilisation des taux d'appels servis de la simulation

Nous supposons que $\tilde{s}_{j,p}$ est le nombre moyen d'agents du centre qui ont traité les appels de type j pour la période considérée. Pour déterminer ce nombre pour chaque type d'appel j, nous utilisons ici les taux d'appels servis par les groupes et leurs staffing pour la période considérée. Soient y_i le nombre d'agents du groupe i et r_{ij} le taux d'appel servi de type j par le groupe i. Le taux global d'appel servi du groupe i est défini par $r_i = \sum_{j \in J} r_{ij}$ où J représente l'ensemble des types que le groupe i peut servir. Nous pouvons ainsi approximer le nombre moyen d'agents du groupe i qui traitent le type j, $\tilde{y}_{ij}$, par

$$\tilde{y}_{ij} = \frac{r_{ij}}{r_i} y_i. \tag{6.5}$$

Nous pouvons ainsi déterminer $\tilde{s}_j$ par la formule suivante.

$$\tilde{s}_{j,p} = \sum_{i \in I} \tilde{y}_{ij}, \tag{6.6}$$

où I est l'ensemble des groupes ayant la compétence pour servir le type d'appel j.

6.2.1.4 QL 4 : utilisation des proportions du temps de service des groupes d'agents par type d'appel

Comme pour la méthode précédente, nous supposons ici aussi que $\tilde{s}_{j,p}$ est le nombre moyen d'agents qui ont eu à traiter les appels de type j pour la période de temps considérée. Nous utilisons cette fois-ci, les proportions des temps de service des groupes d'agent pour chacun des types d'appels pour déterminer $\tilde{s}_{j,p}$. Soit $T_{ij} = \frac{r_{ij}}{\mu_j}$ le temps de service global du type d'appel j par le groupe i où μ_j est le taux de service des appels de type de j. Le temps total de service du groupe i est défini par $T_i = \sum_{j \in J} T_{ij}$ où J représente l'ensemble des types que le groupe i peut servir. Nous pouvons ainsi déterminer le nombre moyen d'agents du groupe i qui ont traité le type j, y_{ij}, par

$$y_{ij} = \frac{T_{ij}}{T_i} y_i. \tag{6.7}$$

Nous pouvons ainsi déterminer $\tilde{s}_j$ comme suit :

$$\tilde{s}_{j,p} = \sum_{i \in I} y_{ij}, \tag{6.8}$$

où I est l'ensemble des groupes ayant la compétence pour servir le type d'appel j.

6.2.1.5 Discussion sur les différentes méthodes

La méthode QL 1 devrait donner les meilleurs résultats parce qu' il utilise l'espérance du temps d'attente conditionnelle à la longueur de la file k, qui est estimée avec des données réellement observées, pour estimer un $\tilde{s}^k_{j,p}$ correspondant à chaque k. La contrainte avec cette méthode est que nous avons besoin d'enregistrer beaucoup d'informations (l'ensemble $\mathcal{D}$), pour chaque type d'appel j à la période p, pour faire les prédictions.

Pour chaque type j, la méthode QL 2 utilise un unique $\tilde{s}$ qui une moyenne agrégée de l'ensemble des $\tilde{s}$ de la méthode QL1. Avec cette méthode, nous avons besoin de sauvegarder une seule information pour chaque type d'appel. Si la variance des $\tilde{s}^k_{j,p}$ est petite dans $\mathcal{D}$ alors la méthode QL2 devrait donner des résultats

sensiblement égaux à ceux QL 1. Par contre, si cette variance est importante, alors QL2 peut donner des performances assez différentes de celles de QL 1.

Les méthodes QL 3 et QL 4 utilisent les mesures de performance observées dans le centre d'appels et le staffing des groupes pour estimer la valeur $\tilde{s}_{j,p}$ pour chaque type d'appel j à la période p. Ces méthodes supposent que le nombre total d'agents dans le système multi-compétence est égal à la somme des agents dans les différents systèmes indépendants, c'est à dire :

$$\sum_{g \in G} s_g = \sum_{j \in J} \tilde{s}_j, \tag{6.9}$$

où G est le nombre de groupes, s_g le staffing du groupe g, et J est le nombre de type d'appels. La véracité de l'équation (6.9) n'est pas prouvée et peut constituer la principale faiblesse de ces méthodes.

6.2.2 Exemple numérique du modèle N sans abandon

Nous considérons un modèle N sans abandons. La journée est constituée d'une seule période de 10 heures. Les processus d'arrivée sont des processus de Poisson de taux constant. Toutes les durées de service sont exponentielles. Nous utilisons la même politique de routage (noté R1) que les modèles N étudiés à la section 4.3.4. Dans nos exemples numériques, pour les appels de type 1, nous supposons que le taux d'arrivée est $\lambda_1 = 30$ appels par heure et le taux de service est $\mu_1 = 2$ appels par heure. Pour le type 2, les paramètres sont respectivement $\lambda_2 = 27$ et $\mu_2 = 3$. Le staffing est de 10 agents pour le groupe 1 et de 15 agents pour le groupe 2.

Les mesures de performance du modèle sont les suivantes. Pour le type d'appel 1, la longueur moyenne de la file d'attente est de 15.6, la probabilité de délai est de 76.3%, le temps d'attente moyen est de 31 minutes (pour tous les clients), et le temps d'attente moyen des clients qui ont attendu avant d'être servis est de 2.2 heures. Pour le type 2, ces mesures sont en moyenne respectivement de : 1.6 clients en moyenne dans la file d'attente, 81.1% de probabilité de délai, 3 minutes d'attente moyenne pour tous les clients, et 33 minutes d'attente moyenne pour les clients qui

ont attendu avant le service. Le partage des compétences est très présent : 65% des appels de type 1 servis ont été répondus par des agents du groupe 1, tandis que les autres 35 % ont été répondus par des agents du groupe 2.

6.2.2.1 Les estimations de $\tilde{s}$ avec les différentes méthodes

Pour estimer $\tilde{s}$ avec chacune des méthodes, nous simulons le modèle pour recueillir toutes les informations nécessaires. La méthode QL1 estime un $\tilde{s}$ qui est conditionnel au nombre k de clients en attente. Le tableau 6.1 donne les valeurs estimées de $\tilde{s}$ pour $k = 0$ à $k = 20$ pour les deux types d'appels. En observant les données, nous constatons que $\tilde{s}$ varie en fonction de k. Pour le type 1, la plus petite valeur de $\tilde{s} = 14.83$ et sa plus grande valeur est $\tilde{s} = 15.93$. Pour le type 2, ses valeurs sont respectivement 12.85 et 14.61. La méthode QL2 qui estime $\tilde{s}$ en utilisant la moyenne pondérée des temps d'attente moyens observés pour les différents k donne une valeur de $\tilde{s} = 15.71$ pour le type 1 et une valeur de $\tilde{s} = 13.39$ pour le type 2.

En simulant le modèle, nous observons les mesures de performances suivantes. Le taux d'appels servis par le groupe 1 pour tous les types est $r_1 = 19.36$. Il est réparti comme suit pour les deux types d'appels. Le taux d'appels servis du type 1 est $r_{11} = 19.36$ et le taux d'appels servis du type 2 est $r_{12} = 0$. Pour le groupe 2, ces taux sont respectivement $r_2 = 37.63$, $r_{21} = 10.65$ et $r_{22} = 26.98$. Sachant que le staffing du groupe 1 est $s_1 = 10$ agents et celui du groupe 2 est $s_2 = 15$ agents, la méthode QL3 donne $\tilde{s} = 14.25$ pour le type d'appel 1 et $\tilde{s} = 10.76$ pour le type d'appels 2. En plus du staffing, la méthode QL 4 qui utilise $\mu_1 = 2$ et $\mu_2 = 3$ donne $\tilde{s} = 15.58$ pour le type 1 et $\tilde{s} = 9.42$ pour le type 2.

La figure 6.2 affiche la courbe de $\tilde{s}$ en fonction de k pour les types 1 avec les différentes méthodes. Elle montre que le $\tilde{s}$ estimé par QL4 est plus proche de celui de QL2 (qui est la moyenne pour des valeurs de QL1) que celui de QL3. Ceci devrait avoir des conséquences sur les performances observées avec les prédicteurs. QL4 devrait donner un RRASE plus petit que QL3 pour le type 1. La figure 6.3 affiche la courbe de $\tilde{s}$ en fonction de k pour les types 2 avec les différentes méthodes.

	$\tilde{s}$ estimé	
k	Type 1	Type2
0	15.9	12.8
1	15.1	13.0
2	14.8	13.2
3	14.8	13.3
4	14.8	13.4
5	14.9	13.6
6	15.0	13.6
7	15.0	13.7
8	15.1	13.7
9	15.2	13.9
10	15.3	13.9
11	15.4	13.9
12	15.4	14.0
13	15.4	14.2
14	15.4	14.3
15	15.5	14.3
16	15.5	14.3
17	15.5	14.3
18	15.6	14.3
19	15.6	14.5
20	15.7	14.6

Tableau 6.1 : Les valeurs de $\tilde{s}$ pour $k = 0$ à $k = 20$ pour QL1

Nous observons que ici le $\tilde{s}$ estimé par QL3 est proche de la moyenne estimée par QL2 que celle estimée par QL4.

6.2.2.2 Performances des prédicteurs

Nous comparons les performances et la précision des prédicteurs LES, QL1, QL2, QL3, QL4, E-LES, et AvgC-LES dans cet exemple. Nous avons effectué des simulations indépendantes et évalué le RRASE des différents prédicteurs. Le tableau 6.2 rapporte les RRASE×100. Il montre que les prédicteurs QL réadaptés sont les plus précis. Ils sont suivis par AvgC-LES et E-LES respectivement. LES est le prédicteur le moins précis.

En comparant les performances des prédicteurs QL, nous constatons que QL1 et

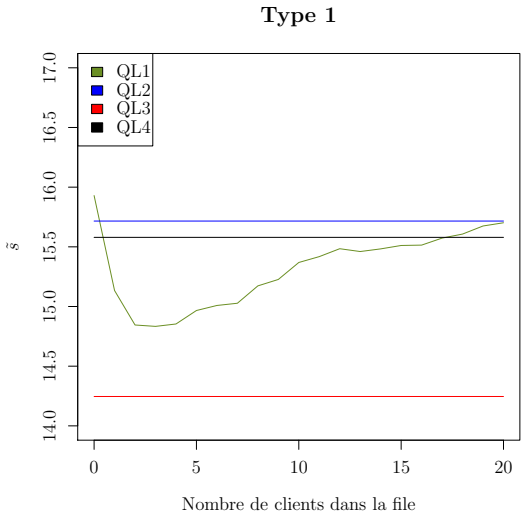

Figure 6.2 : Variation de $\tilde{s}$ en fonction k pour le type d'appel 1.

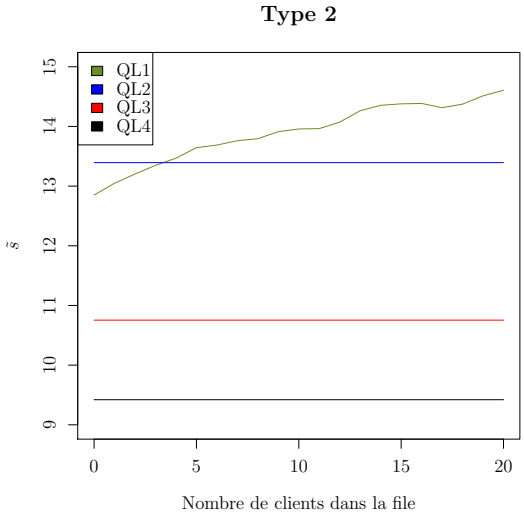

Figure 6.3 : Variation de $\tilde{s}$ en fonction k pour le type d'appel 2.

QL2 donnent sensiblement les mêmes résultats, et sont les plus précis. Cependant, nous constatons que les performances de QL3 et QL4 diffèrent selon le type d'appels. Pour le type 1, QL4 donne presque les mêmes résultats que QL1 et QL2, mais pour le type 2, il est de loin moins précis. QL3 est moins que QL4 pour le type 1, par contre il est plus précis que ce dernier pour type 2.

Au moment où nous écrivons ces lignes, nous n'avons pas encore compris pourquoi un tel comportement des prédicteurs. Dans le futur, nous allons essayer de comprendre le pourquoi. Le fait que les prédicteurs QL sont plus précis que E-LES et AvgC-LES est encourageant et motive à continuer la recherche.

	LES	QL1	QL2	QL3	QL4	E-LES	AvgC-LES
Type 1	45.6	29.4	29.4	33.4	29.5	41.9	35.8
Type 2	77.8	50.6	50.7	53.6	55.3	74.6	58.0

Tableau 6.2 : RRASE×100 pour le modèle N sans abandon avec une seule période.

6.2.2.3 Robustesse des prédicteurs QL adaptés face à la variation des taux d'arrivée

Dans les centres d'appels, les paramètres réellement observés sont la plupart du temps différents de ceux prédits. Dans cette section, nous allons observer le comportement des prédicteurs QL1 face à la variation des taux d'arrivée.

Nous avons augmenté les taux d'arrivés pour les deux types d'appels (λ_1 et λ_2) simultanément de 1 et 2% . Nous constatons à chaque fois une augmentation significative de la longueur moyenne de la file d'attente pour le type 1 et une faible augmentation pour le type 2. La probabilité de délai augmente à chaque fois d'environ 4% pour les deux types d'appels.

Les tableaux 6.3, et 6.4 donnent respectivement les RRASEs×100 des prédicteurs dans les deux cas. Nous rappelons que les prédictions sont effectuées en utilisant les $\tilde{s}$ estimés avec les données du modèle initial (modèle avant augmentation des taux d'arrivée). Nous constatons que les nouveaux prédicteurs sont toujours plus précis que LES, E-LES, et AvgC-LES.

	LES	QL1	QL2	QL3	QL4	E-LES	AvgC-LES
Type 1	50.3	34.5	32.3	36.1	32.4	48.7	39.7
Type 2	78.9	51.8	51.6	54.3	56.3	75.1	59.3

Tableau 6.3 : RRASE modèle N avec une augmentation de 1%

	Prédicteurs						
	LES	QL1	QL2	QL3	QL4	E-LES	AvgC-LES
Type 1	58.4	37.3	35.2	39.6	35.2	55.4	47.3
Type 2	79.6	55.6	55.4	58.3	60.2	76.3	60.4

Tableau 6.4 : RRASE modèle N avec une augmentation de 2%

6.2.3 Variation de $\tilde{s}$ en fonction du routage

Pour montrer que $\tilde{s}$ varie en fonction de la politique de routage, nous allons maintenant utiliser le même modèle N avec les mêmes paramètres, mais cette fois-ci avec une politique différente de R1. Nous utilisons une politique de routage (R2) pour laquelle il n'y a pas de priorité pour les agents du groupe 2. Elle fonctionne comme suit. Les agents du groupe 2 donnent la même priorité aux deux types. Si un agent du groupe 2 devient disponible et qu'il y a des clients en attente dans les deux files, la priorité sera accordée à l'appel qui a attendu le plus longtemps. Les appels du même type sont de premier arrivé, premier servi. Pour un appel de type 1, s'il y a un agent libre dans chacun des deux groupes, le routeur va l'affecter à l'agent qui est resté le plus longtemps sans servir d'appel. La figure 6.4 montre la variation de $\tilde{s}$ avec les politiques des routages R1 et R2 pour les types 1 et 2. Nous constatons dans les deux cas que la distribution de $\tilde{s}$ est différente pour chaque politique de routage.

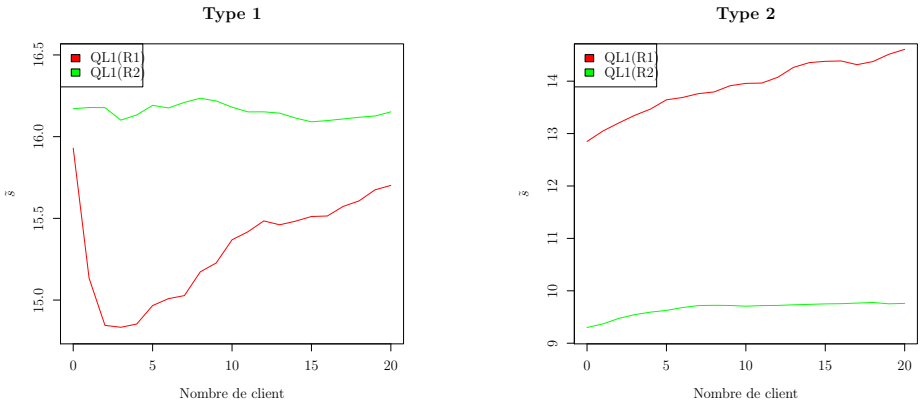

Figure 6.4 : QL1, Variation de $\tilde{s}$ en fonction du routage pour $k = 0$ à $k = 20$.

6.3 Prédiction de la distribution conditionnelle des temps d'attente

Les prédicteurs de délai proposés dans cette thèse comme la plupart des prédicteurs dans la littérature estiment une espérance conditionnelle du temps d'attente prévu pour le client. Cette espérance est annoncée au client comme étant son temps d'attente estimé. Il est rare que le client observe exactement le temps d'attente annoncé. Le temps observé est soit plus petit ou plus grand que celui annoncé. Lorsque le délai attendu devient supérieur au délai annoncé, le client peut devenir très impatient (Mowen et al., 1993). Ceci peut entraîner une augmenter des abandons dans le système et diminuer la satisfaction du client vis-à-vis du fournisseur de service. Pour éviter cela, nous pensons qu'il est préférable de donner plus d'information au client à propos de son délai d'attente comme par exemple une estimation de la distribution conditionnelle de son temps d'attente ou au moins certains quantiles de cette distribution.

Nous savons que pour le modèle M/M/s (un processus de Poisson pour les arrivées, les temps de service exponentiels de moyenne μ^{-1}), et s serveurs) à chaque fois que tous les serveurs sont occupés, le temps jusqu'à la prochaine fin de service est une exponentielle de moyenne de $1/s\mu$, indépendamment du passé. Par conséquent, le temps d'attente avant le démarrage du service pour une nouvelle arrivée avec $s + k$ clients dans le système est la somme de $k + 1$ variables aléatoires exponentielles i.i.d de moyenne de $1/s\mu$ chacune, qui suit une distribution d'Erlang de forme $k+1$ et d'intensité $s\mu$. Pour ce modèle, nous pouvons informer les clients non seulement de la moyenne de son attente, mais aussi de l'histogramme ou d'un graphique de sa densité de probabilité.

Cependant, pour la plupart des modèles de file d'attente simples, la distribution des temps d'attente d'un client ayant trouvé d'autres clients dans la file est inconnue et n'est pas facile à déterminer. Elle devient encore plus difficile pour les système multi-compétences. Nous pensons qu'il serait intéressant de développer des méthodes qui permettent son estimation afin de fournir aux clients des plus d'information.

La première idée que nous avons est la suivante : Au chapitre 4, nous avons développé des prédicteurs qui utilisent un ensemble de données $\mathcal{D} = \{\boldsymbol{x}, y\}$ où $\boldsymbol{x}$ est un vecteur qui définit l'état du système et y le temps d'attente observé pour un client qui a observé cet état à son arrivée. Nous pouvons regrouper ces données autrement. Pour chaque état $\boldsymbol{x}$ du système, nous collectons un ensemble $\boldsymbol{y}^T$ où $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)$ avec y_i une réalisation du temps d'attente pour un client qui observée l'état $\boldsymbol{x}$ à son arrivée. Nous pouvons chercher la distribution qui fit au mieux les observons de $\boldsymbol{y}$ et estimer les paramètres de $\boldsymbol{y}$ pour chaque $\boldsymbol{x}$. Nous pouvons par la suite construire un ensemble de données $\mathcal{D}' = \{\boldsymbol{x}, \boldsymbol{p}^T\}$ où le vecteur de paramètres. Nous pouvons par exemple utiliser les réseaux de neurones pour les prédire les paramètres d'un $\boldsymbol{x}$ jamais observé dans nos données.

CHAPITRE 7

MODÉLISATION DES DURÉES DE SERVICE DANS LES CENTRES D'APPELS

7.1 Introduction

Dans ce chapitre, nous présentons notre contribution dans l'article Ibrahim et al. (2016b) sur la modélisation des durées de services dans les centres d'appels multi-compétences. Cet article propose une modélisation réaliste des temps de service dans les centres d'appels. En particulier, il réalise, à grande échelle et en profondeur, une investigation empirique des temps de service dans les centres d'appels. Une analyse des données recueillies au centre d'appels d'Hydro-Québec (HQ) y est effectuée. Le centre d'appels réel est complexe et composé de plusieurs types d'appels distincts et de nombreux agents hétérogènes. Les données montrent que les temps de service diffèrent considérablement entre ces agents, varient dans le temps, et présentent une forte corrélation sérielle et croisée. Des modèles pour les temps de service qui en prennent en compte toutes ces caractéristiques, et qui s'ajustent bien aux données réelles ont été proposés.

Il était important de vérifier que ces modèles sont des outils fiables pour prédire les moyennes des temps de service des agents mieux que les modèles de références ("*benchmark*") considérés, mais surtout montrer par la simulation que cette meilleure prédiction aura un impact sur les performances du système. La simulation est un outil important qui peut être utilisé pour évaluer les mesures de performance telles que le niveau de service et le temps d'attente moyen, et pour construire des horaires de travail pour les agents et des règles de routage par des algorithmes d'optimisation stochastique (Avramidis et al., 2010, Chan et al., 2014).

Nous avions à notre disposition un bon simulateur des centres d'appels Contact-Centers (Buist, 2009, Buist et L'Ecuyer, 2005). Ce simulateur a été développé à l'université de Montréal par Eric Buist dans la cadre de sa thèse sous la supervi-

sion du professeur Pierre L'Ecuyer. Concernant la distribution des temps de service des agents, le simulateur offrait seulement l'option de spécifier une distribution du temps de service par groupe d'agent et par type d'appel, mais n'offrait pas l'option de spécifier une distribution du temps de service par agent et par type d'appels. Cette dernière option était nécessaire pour simuler un centre d'appels avec les modèles développés dans l'article. Mon premier objectif etait : (i) de développer un module permettant de spécifier une distribution du temps de service distinct pour chaque agent et pour chaque type d'appel dont il possède les compétences et (ii) d'intégrer ce module dans le simulateur. Notre second objectif était de montrer par des exemples tirés de nos données (données recueillies au centre d'appels de HQ) que les nouveaux modèles proposés prédisent mieux la moyenne des temps de service que les modèles de benchmark et par la suite montrer par des simulations que les mesures de performances du centre d'appels obtenues avec les différents modèles (les benchmark et les nouveaux modèles) peuvent être très différentes. Dans ce chapitre nous présenterons le travail et les résultats obtenus dans ce second objectif.

Le reste du chapitre sera organisé comme suit. Dans la section 7.2, nous faisons une revue de la littérature des travaux sur la modélisation des durées de service. La section 7.3 présente l'analyse préliminaire faite sur les données. À la section 7.4, nous faisons une description des nouveaux modèles proposés. La section 7.5 décrit la qualité de l'ajustement des modèles aux données. Nous présenterons à la section 7.6 les résultats des prédictions et à la section 7.7 les résultats des simulations. Nous terminerons ce chapitre par une conclusion et des remarques à la section 7.8.

7.2 Revue de littérature

Traditionnellement, les chercheurs et les praticiens ont utilisé les modèles de file d'attente d'Erlang standard pour analyser les opérations dans les centres d'appels. Dans les modèles de files d'attente d'Erlang, les temps de service des agents sont modélisés comme des variables aléatoires indépendantes et identiquement distribuées

exponentielles avec une moyenne constante. Au-delà de cette hypothèse standard de modélisation, il y a des conséquences opérationnelles importantes, comme en témoignent les multiples avancées dans la littérature récente.

7.2.1 Hétérogénéité des agents

Il existe plusieurs articles qui étudient des modèles de files d'attente avec des serveurs hétérogènes, avec des applications pour la gestion des centres d'appels. Une question centrale qui se pose dans ce contexte est de savoir comment router les appels entrants vers des agents hétérogènes de manière à minimiser une mesure de performance donnée, par exemple le temps d'attente moyenne. Compte tenu de la complexité de ce problème, la plupart des articles ont recours à l'adoption de politiques de routage optimales dans les systèmes à grande échelle dans des conditions à trafic intense ; voir par exemple Armony (2005), Armony et Mandelbaum (2011), Armony et Ward (2010), Gurvich et Whitt (2009) et les références citées. Mehrotra et al. (2012) ont recours à une étude numérique pour caractériser les performances globales en utilisant le temps d'attente moyen des clients et le taux de service global. En général, ces articles montrent que les décisions de contrôle peuvent réellement bénéficier de l'hétérogénéité des agents, par exemple, router des appels entrants vers les agents libres les plus rapides réduit le temps d'attente des clients.

Il y a très peu de recherches empiriques pour étayer ces travaux théoriques. Au meilleur de notre connaissance, la seule exception est Gans et al. (2010) qui ont analysé les données d'un centre d'appels et identifié à la fois à court terme et à long terme des facteurs associés à l'hétérogénéité des agents en pratique. Ils ont également décrit les résultats d'une petite étude de simulation illustrant les conséquences opérationnelles d'ignorer une telle hétérogénéité. Gans et al. (2010) ont indiqué qu'une extension intéressante de leurs travaux est d'incorporer les effets aléatoires dans les modèles du temps de service afin "de capturer la dépendance au sein des appels traités par le même agent, et de permettre la compréhension de la population d'agents dans son ensemble" (p. 118). Ces effets aléatoires ont été

considérés dans les modèles proposés dans ce chapitre et qui sont présentés dans la section suivante.

7.2.2 Dépendances entre les temps de service

Les temps de service en pratique sont souvent dépendants. Pour un exemple, un agent peut être surchargé de travail dans des périodes données (par exemple, dans les périodes de congestion) et cela pourrait affecter ces performances dans tous les services qu'il effectue pendant ces périodes de travail, résultant généralement à ce que l'agent soit lent ou rapide ; voir Delasay et al. (2016), Dong et al. (2015), Feldman et al. (2015) et les références citées. Dans ce cas, les agents (serveurs) peuvent être considérés comme des décideurs stratégiques qui influencent leur propre taux de service. La conséquence d'un tel comportement stratégique est que les temps de service successifs sont dépendants. Pour un deuxième exemple, dans un centre d'appels technique, il peut y avoir un défaut de produit, raison pour laquelle il y a de multiples appels connexes, dont les durées sont toutes plus que la moyenne. Dans cet exemple aussi, les temps de service (durées des appels) sont dépendants. Pour un troisième exemple, dans un centre d'appel d'urgence, plusieurs appels entrants pourraient être liés un même incident médical, auquel cas les durées de ces appels seraient aussi bien dépendants.

Il y a une théorie bien développée sur l'étude de l'impact sur les performances de la dépendance entre les temps de service dans les systèmes de files d'attente à serveur unique ; par exemple, voir le chapitre 9 de Whitt (2002) pour un traitement détaillé. Cependant, Dong et al. (2012) sont parmi les premiers à considérer le cas multi-serveur, ce qui est plus raisonnable d'un point de vue pratique. Ils ont considéré une séquence stationnaire de temps de service faiblement dépendants et ont démontré que, dans la limite de trafic intense, l'impact de ces dépendances est déterminé par la fonction de répartition bivariée des temps de service. Dans leur étude numérique, ils ont considéré une séquence EARMA ("*exponential autoregressive-moving average*") de temps de service, qui est stationnaire avec des distributions marginales exponentielles, et la structure de corrélation d'un processus autorégres-

sif de moyenne mobile. Ces auteurs ont démontré, par une analyse théorique et des simulations, comment les dépendances entre les temps de service peuvent modifier significativement les performances des grands systèmes. En particulier, ils ont montré que ces corrélations influent fortement la distribution du nombre de clients dans la file d'attente qui, à son tour, affects les décisions du staffing. Dong et al. (2012) ont conclu leur papier en appelant à "des études empiriques pour estimer la grandeur de la dépendance des temps de service dans les applications" (p. 278). Une telle étude est conduite dans les modèles présentés dans la section suivante.

7.2.3 Dépendances avec le temps

Il y a relativement peu de travaux qui considèrent des modèles file d'attente avec des taux de service variant dans le temps, puisque cette fonctionnalité complique sensiblement l'analyse. Certaines exceptions sont Aldor-Noiman et al. (2009), Liu et Whitt (2011), Mandelbaum et al. (1999) et les références qui y figurent. Ces articles démontrent l'impact opérationnel d'inclure des taux de service variant dans le temps ; leurs résultats appliquent une forme générale et ne supposent pas une forme spécifique pour la dépendance du temps dans le taux de service. Aldor-Noiman et al. (2009) ont utilisé les prévisions des futurs nombres d'arrivée et moyennes des temps de service pour estimer les charges futures dans les centres d'appels. Aldor-Noiman et al. ont autorisé les moyennes des temps de service d'être dépendantes du temps, et ont montré comment les erreurs dans la prédiction des charges futures peuvent influer sur les décisions de staffing. Leur article suppose des agents homogènes et un seul type d'appel. Les modèles de temps de service considérés ici sont fonction du temps, mais dans un cadre beaucoup plus complexe, avec des types d'appels multiples et de nombreux agents hétérogènes.

7.2.4 Distribution Log-normale

Brown et al. (2005) ont réalisé une analyse statistique détaillée des données d'un centre d'appels et ont montré que les temps de service ne sont pas exponentiellement

distribués, comme on l'a traditionnellement supposé, et que la loi log-normale est mieux adaptée pour la distribution du temps de service à sa place. Deslauriers (2003), Pichitlamken et al. (2003) ont également observé la même chose. Motivé par ceci, Shen et Brown (2006) ont proposé une nouvelle méthode pour l'inférence sur les courbes de régression non paramétriques lorsque les erreurs sont distribuées selon la loi log-normale. Ils ont illustré leur méthode à la fois par une étude par simulation et l'analyse des données d'un centre d'appels de la vie réelle. Mandelbaum et Zeltyn (2011) préconisent un processus des temps de service qui est modélisé comme un processus de Markov à temps continu absorbant à état fini. Ici, même si nous utilisons les informations supplémentaires lors de la modélisation des temps de service, comme le temps post-traitement de l'appel, nous continuons d'assumer la log-normalité des temps de service individuels.

7.3 Analyse préliminaire des données

Les données utilisées ici ont été recueillies au centre d'appels de HQ sur la durée d'un an, allant du 3 janvier 2011 au 31 décembre 2011. Le centre d'appels est virtuel avec plus de 15 emplacements à travers le Québec, et est ouvert du lundi au vendredi et est fermé le week-end (samedi et dimanche). Les données sont constituées des moyennes quotidiennes des temps de service pour plusieurs agents et différents types d'appels. Même s'il est souhaitable d'étudier les données d'appel par appel, de nombreux centres d'appels recueillent encore régulièrement des données agrégées à la place ; voir par exemple Oreshkin et al. (2016), Pinedo et al. (1999). Par conséquent, il est important de développer des modèles de temps de service dont les paramètres peuvent être estimés avec ces données agrégées, comme nous le faisons ici. En plus des moyennes quotidiennes des temps de service, les données contiennent des informations sur le nombre quotidien d'appels traités par chaque agent, par type d'appel. Les types d'appels se distinguent à la fois par la nature de la demande de service et la langue, soit en français ou en anglais, dans laquelle l'appel est traité.

Un temps de service est souvent constitué d'une première partie, assurée par un serveur vocal interactif (IVR), et une seconde partie où l'appel est traité par un agent. Puisque nous sommes intéressés à la modélisation des temps de service du point de vue des agents, nous ne considérons pas la partie IVR parce que les agents ne sont pas nécessaires pour cette partie. Le temps passé par les clients dans l'IVR est par exemple étudié par Colladon et al. (2013), Salcedo-Sanz et al. (2010). Du point de vue d'un agent (notre point de vue), un temps de service individuel est la somme de : (i) le temps passé effectivement à parler au client (temps d'appel), et (ii) le temps de post-appel passé par l'agent à noter les questions liées à l'appel, au cours de laquelle il reste indisponible.

7.3.1 Vue d'ensemble

Dans notre ensemble de données, il y a 148 types d'appels traités par un groupe de 1 655 agents. La plupart des agents ont des compétences différentes et ils traitent différents types d'appels en fonction de ces compétences. Au total, il y a 16 328 combinaisons distinctes agent / type d'appel, où chaque combinaison correspond à un traitement d'un type d'appel particulier par un agent. De nombreux types d'appels ont très peu d'appels correspondants, et il n'est pas intéressant pour nous de les étudier. Nous enlevons tous les types d'appels qui ont moins de 10 appels au total, à travers tous les agents de nos données, et on s'est retrouvé avec 86 types d'appels traités par un total de 1 562 agents.

Nous traçons à la figure 7.1 la moyenne du nombre d'agents qui répondent aux appels par jour de semaine, avec des intervalles de confiance de 95 pour-cent qui correspondent aux quantiles empiriques 2.5 pour cent et 97.5 pour-cent, basés sur les agents qui ont traité au moins 10 appels dans les données. Nous voyons que le nombre d'agents est grandement variable les lundis, et que les vendredis ont moins d'agents, en moyenne. Dans la Figure 7.2, nous traçons la moyenne du volume total d'appels par jour de semaine, en incluant tous les types d'appels. La Figure 7.2 montre que les volumes d'appels des lundis présentent la plus grande variance, et que les volumes d'appels des vendredis sont les plus bas en moyenne.

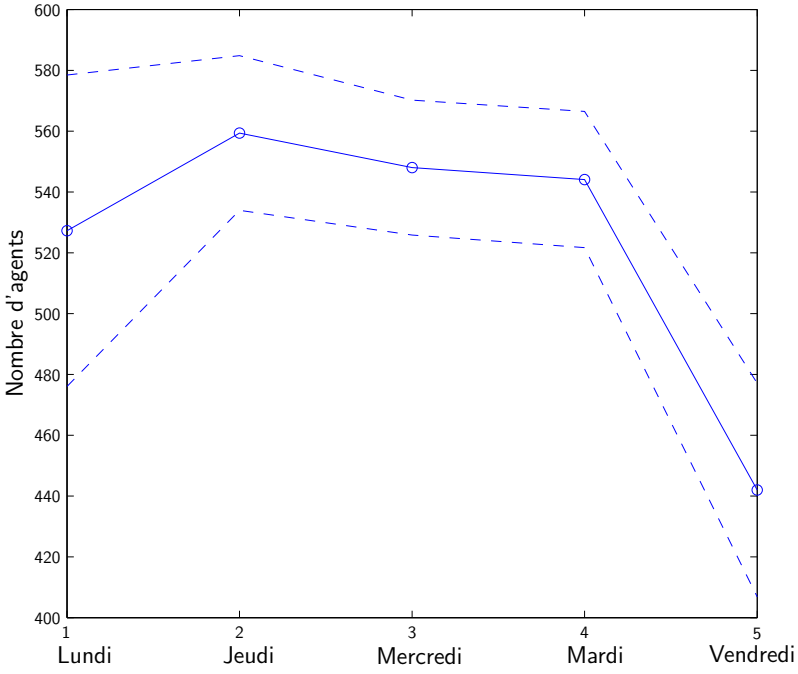

Figure 7.1 : Nombre moyen d'agents par semaine et les bandes de confiance à 95% correspondant.

Les agents traitent généralement plus d'un type d'appel pour un jour donné ; aussi chaque type d'appel est généralement traité par plus d'un agent. Par exemple, environ 400 agents traitent 1 à 3 types d'appels différents, et environ 25 types d'appels sont traités par environ 65 agents chacun. La médiane du nombre total de types d'appels traités par agent (au cours de la période d'un an) est de 13, et la médiane du nombre d'agents traitant un type d'appel est donnée 33.

7.3.2 Statistiques sur les temps de service

Nous rapportons plusieurs observations empiriques importantes de nos données. Nos modèles stochastiques pour les temps de service sont développés pour incorporer de telles caractéristiques.

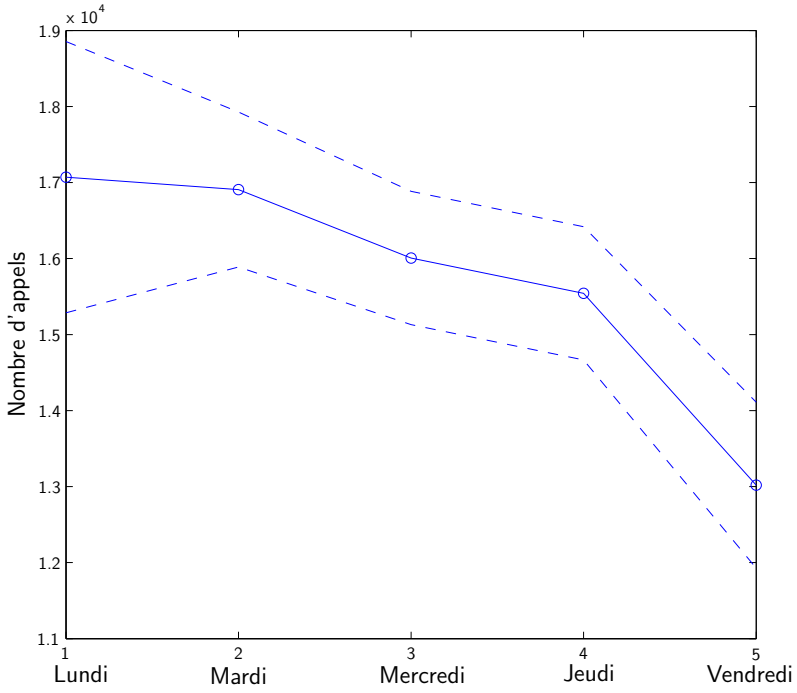

Figure 7.2 : Nombre moyen d'appels répondus et les bandes de confiance à 95% correspondant.

7.3.2.1 Variations à travers les types d'appel

La figure 7.3 donne un nuage de points des moyennes empiriques et variances des temps de service pour les différents types d'appels de nos données. Chaque point correspond à une paire (moyenne, variance), correspondant à un type d'appel donné. La figure 7.3 montre qu'il existe des différences significatives dans les moyennes et les variances à travers différents types d'appels. Comme prévu, la figure 7.3 montre que les types d'appels avec des durées plus longues présentent généralement une variance plus élevée. Nous prenons en compte cette variation entre les types d'appels dans les nouveaux modèles qui sont proposés à la section 7.4.

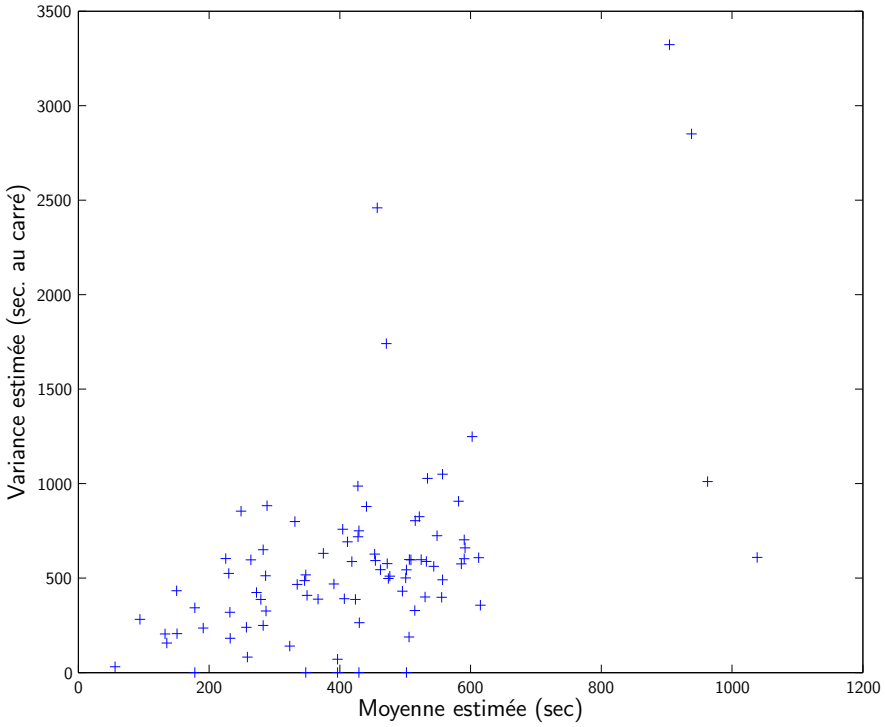

Figure 7.3 : Chaque point correspond à une paire (moyenne, variance) pour type donné.

7.3.2.2 Hétérogénéité des agents

Les distributions des temps de service pour le même type d'appel varient considérablement selon l'agent. Dans les figures 7.4 et 7.5, nous illustrons cette hétérogénéité des agents. Nous traçons la moyenne des temps de service pour deux types d'appels : A, qui est traité par 991 agents, et B, qui est traité par 997 agents, en fonction du nombre total d'appels répondus (sur la période d'un an couverte par nos données) par chaque agent.

La ligne horizontale dans chaque figure indique le temps de service moyen global pour tous les agents, pour chaque type d'appel. Les figures 7.4 et 7.5 montrent qu'il existe une variabilité importante dans les temps de service à travers tous les agents. Les figures 7.4 et 7.5 montrent également qu'il y a clairement des "groupes" d'agents qui semblent se comporter d'une manière à peu près similaire (en ayant des moyennes soit plus courtes ou plus longues que la moyenne générale des temps de service). En général, les agents qui ont traité de nombreux appels au cours de l'année sont beaucoup plus rapides en moyenne que ceux qui n'ont manipulé que quelques appels. Ces derniers sont soit des agents qui ont traité très peu d'appels en général, ou ceux qui ont la plupart du temps traité d'autres types d'appels. En général, il semble que les agents qui ont traité plus d'appels ont tendance à présenter moins de variances dans leur temps de service. En d'autres termes, la plus grande dispersion est principalement donnée par les agents qui sont moins expérimentés (ceux qui ont répondu à moins d'appels).

Dans les figures 7.6 et 7.7, nous traçons des estimations des variances des temps de service pour tous les agents ayant traité les appels de type A et B, respectivement, en fonction du nombre total d'appels de ce type répondu par l'agent. Les figures 7.6 et 7.7 confirment qu'il existe des différences claires dans la variance du temps de service à travers les agents.

Dans la figure 7.8, nous traçons les temps de service moyens de quatre agents ayant traité des appels de type B, en fonction du temps (indice de la journée). De plus, nous incluons des lignes horizontales correspondant au temps de service

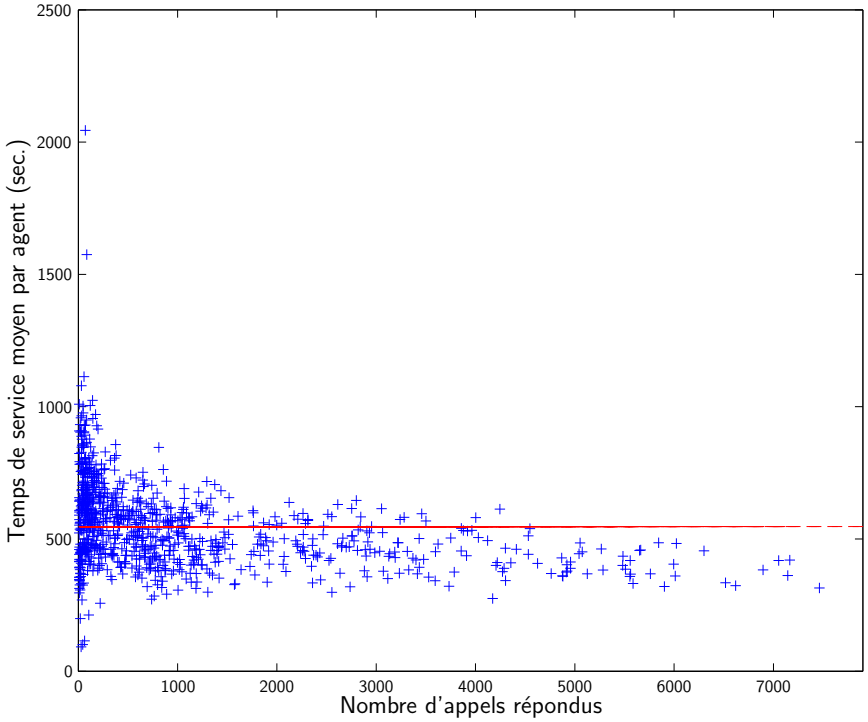

Figure 7.4 : Moyenne du temps de service pour différents agents traitant le type d'appel A en fonction du nombre total d'appels répondus par année. La ligne horizontale est la moyenne globale pour tous les agents.

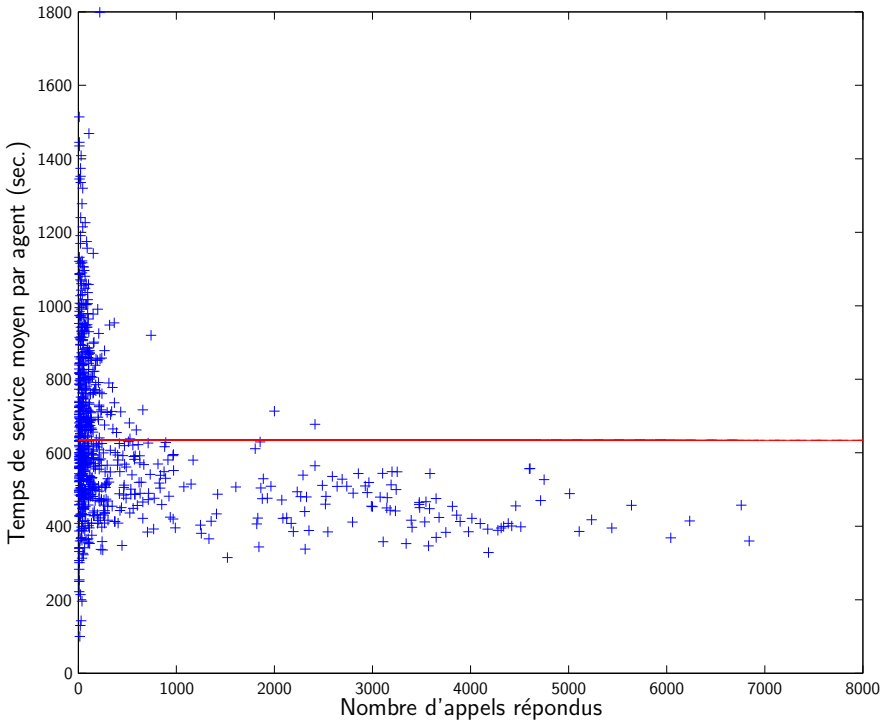

Figure 7.5 : Moyenne du temps de service pour différents agents traitant le type d'appel B en fonction du nombre total d'appels répondus par année. La ligne horizontale est la moyenne à travers tous les agents.

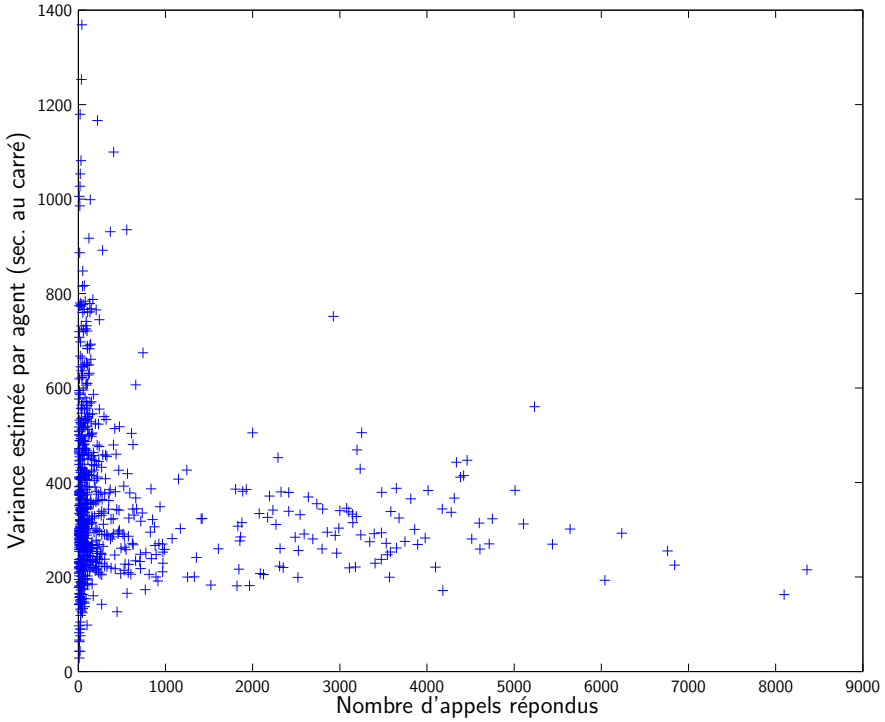

Figure 7.6 : Les variances du temps de service estimées pour les agents traitant le type d'appel A en fonction du nombre total d'appels répondus par année.

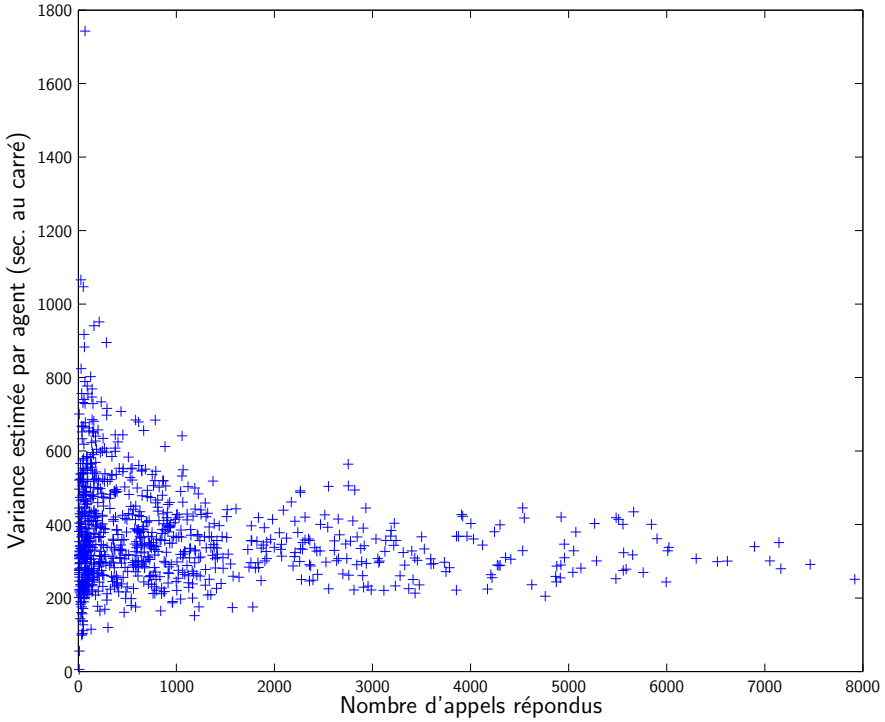

Figure 7.7 : Les variances du temps de service estimées pour les agents traitant le type d'appel B en fonction du nombre total d'appels répondus par année.

moyenne globale pour ces agents. La figure 7.8 montre que différents agents ont un comportement différent. En effet, les deux agents en haut sont évidemment plus lents que les deux agents en bas, et leurs temps de service ont aussi une variance plus élevée.

7.3.2.3 Dépendance du temps

En plus de variabilité entre les différents agents, nos données montrent que le temps de service moyen pour un agent donné et un type d'appel donné varie considérablement au fil du temps.

Dans la figure 7.9, nous traçons les temps de service moyens quotidiens pour un agent traitant quatre types d'appels différents, en fonction du temps. Ces moyennes quotidiennes varient clairement avec le temps. La figure 7.9 illustre un phénomène qui pourrait être important d'un point de vue opérationnel : l'agent semble ralentir quand il gère plusieurs types d'appels. En effet, ce fait est apparent au jour 208 lorsque l'agent commence le traitement des appels de type 4. Par la suite, la figure 7.9 montre que les temps de service moyens des types d'appels 1 et 2 augmentent. Sur la base de ces observations, nous avons expérimenté avec notamment le nombre de types d'appels traités par un agent comme variable dans notre modèle du temps de service. Nous n'avons pas inclus ces modèles dans le présent document, car ils conduisent à des prédictions moins précises de la moyenne des temps de service dans l'échantillon de test de nos données pour tous les agents. Peut-être que cela pourrait être utilisé seulement pour certains agents pour lesquels nous disposons de suffisamment de données avant et après le changement. La figure 7.8 illustre également que les temps de service moyens fluctuent au fil des jours successifs.

Dans la figure 7.10, nous illustrons la dépendance par rapport au temps en traçant l'évolution, au fil du temps des temps de service moyens quotidiens pour un agent a_1 qui traite les appels de type A . Dans la figure 7.10, nous incluons aussi le meilleur ajustement linéaire pour les données. Ce tracé montre clairement une tendance à la hausse dans les temps de service moyens pour cet agent. Dans nos données, nous avons observé des tendances vers le haut et vers le bas, selon

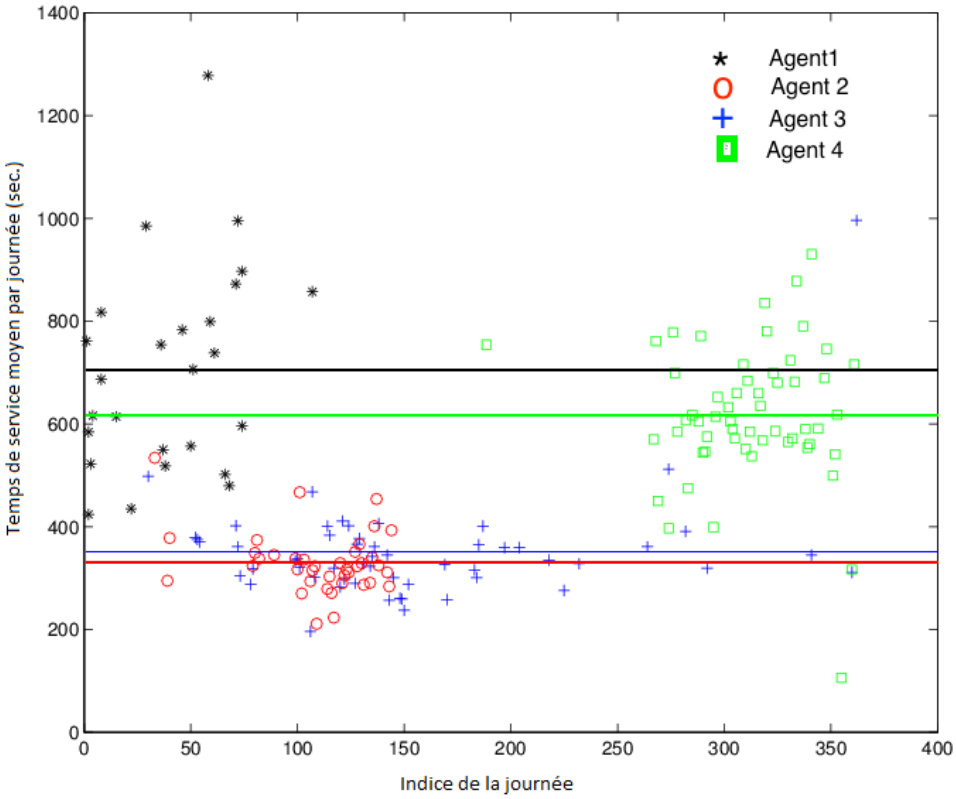

Figure 7.8 : La moyenne des temps de service pour 4 agents traitant le type d'appels B versus indice de la journée.

l'agent. Une explication de tendances à la baisse est que les agents apprennent avec le temps ; voir Gans et al. (2010) pour plus de soutien empirique. Il peut y avoir beaucoup d'autres explications pour de telles tendances. Par exemple, avec des tendances à la hausse, il se peut que les agents se lassent et deviennent moins motivés à répondre rapidement aux appels.

7.3.3 La cohorte C de 200 agents

Le nombre total d'appels traités par agent varie considérablement entre les agents dans nos données. Le maximum est de 14 715 appels traités pour un agent au cours de la période d'un an, mais des centaines d'agents ont répondu à très peu d'appels. Pour ces agents, il est difficile d'ajuster les modèles de temps de

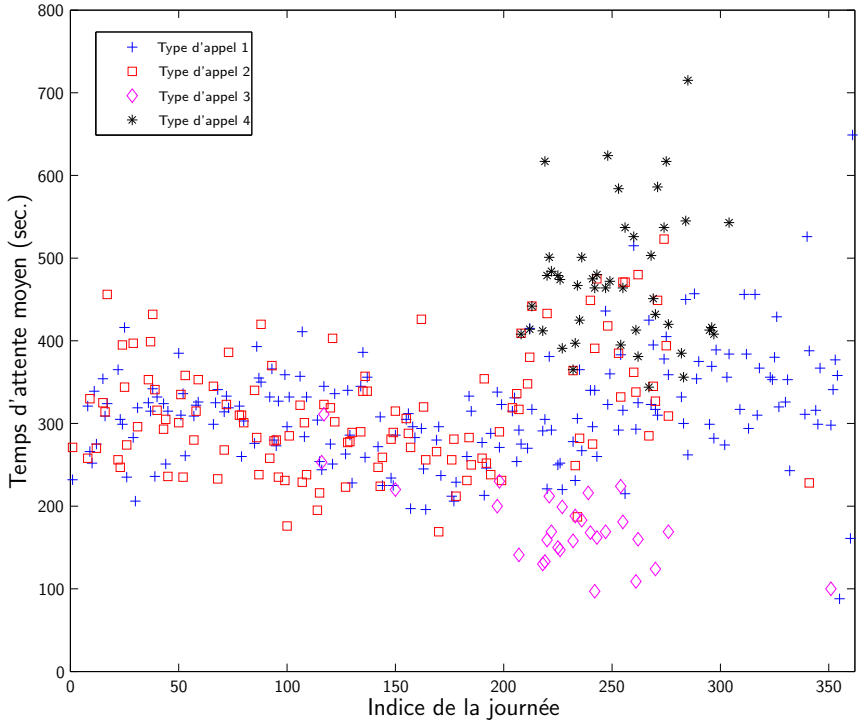

Figure 7.9 : La moyenne journalière des temps de service pour un agent traitant de multiples types d'appels et dont la liste des compétences augmente au jour 208.

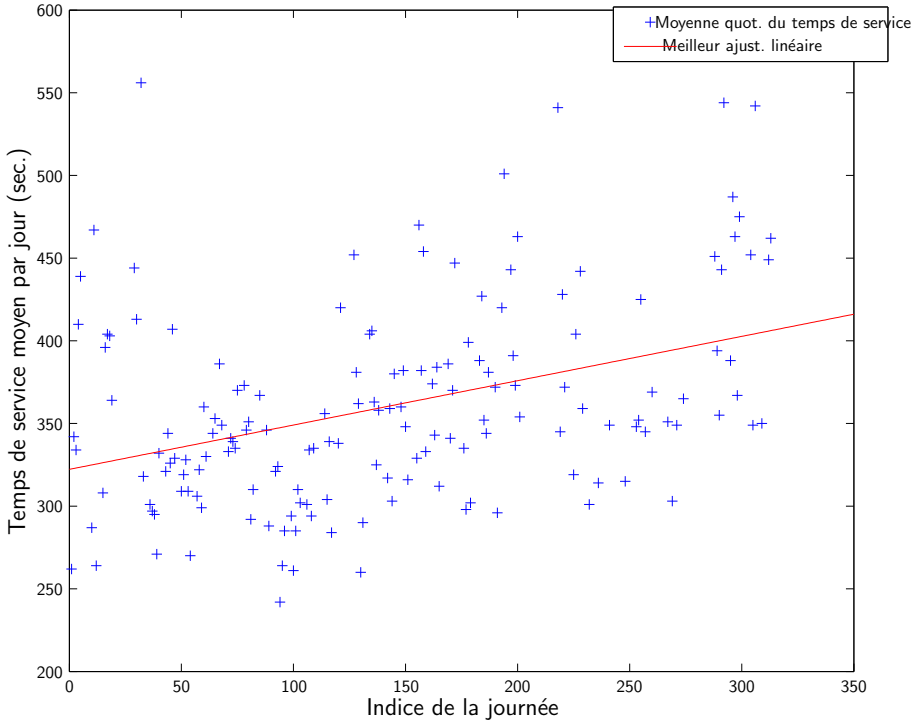

Figure 7.10 : L'évolution de la moyenne des temps de service de l'agent a_1 pour le type d'appels A, et le meilleur ajustement linéaire.

service et faire des prédictions fiables. En outre, avec l'insuffisance des données, il est difficile de parvenir à des résultats significatifs. Pour le reste de cette étude, nous limitons notre attention à des agents qui ont répondu un nombre relativement important d'appels ; en particulier les 200 agents qui ont répondu le plus d'appels au cours de l'année. Ces 200 agents ont répondu à un total de 1 175 178 appels, ce qui correspond à environ la moitié du nombre total des appels entrants au centre au cours de l'année. Pour chacun de ces 200 agents, nous avons supprimé les paires (agent, type d'appel) où l'agent a traité moins de10 appels par jours dans notre ensemble de données. Nous faisons cela pour éviter de considérer les paires (agent, type d'appel) avec trop peu d'observations.

Il y a un total de 550 paires différentes (agent, type d'appel) qui restent dans notre cohorte. Dans le reste du chapitre, nous nous référons à notre cohorte d'agents comme étant la cohorte C. Il est important de noter que ce ne sont pas les 200 agents les plus à droite des figures 7.4-7.7. Au total, ces 200 agents traitent 30 différents types d'appels, et le nombre de compétences par agent varie de 1 à 8. Le nombre moyen de compétences par agent, dans ce sous-ensemble des données, est de 3.9.

7.4 Modèles de Temps de Service

Dans cette section, nous proposons des modèles alternatifs pour les processus du temps de service. Nous commençons par décrire deux modèles de référence ("*benchmark*") qui imitent la pratique standard.

7.4.1 Les Modèles de benchmark B1 et B2

L'analyse préliminaire de la section 7.3 suggère que les temps de service dépendent fortement de l'agent et du type d'appel considéré ; voir les Figures 7.4-7.10. Soit $S_{i,j}$ le temps de service d'un appel de type j traité par l'agent i où $j = 1, 2, \ldots, J$ et $i = 1, 2, \ldots, I$.

Dans le premier modèle de référence, Modèle B1, nous supposons que les $S_{i,j}$ sont des variables aléatoires i.i.d d'une distribution log-normale d'espérance m_j et

de variance v_j, pour chaque i et j, où m_j et v_j dépendent uniquement du type d'appel j. Dans notre deuxième modèle de référence, Modèle B2, nous supposons que la valeur espérée est $m_{i,j}$ et sa variance $v_{i,j}$ dépendent à la fois du type d'appel j et l'agent i.

Puisque les données sont en général constituées que de moyennes journalières agrégées de temps de service, au lieu des données détaillées appel par appel, il n'est pas immédiatement visible de voir comment calculer des estimations ponctuelles pour les espérances et les variances. Pour ce faire, nous adoptons ici la méthode des moments comme dans Deslauriers (2003). Sinon, pour une revue des méthodes d'estimation (pour la moyenne) avec des données plus détaillées, voir Shen et Brown (2006).

Méthode des Moments. Nous fournissons des détails supplémentaires pour cette méthode en nous concentrant sur l'estimation pour le modèle B2. Pour le modèle B1, nous faisons la même chose, mais ne nous faisons pas de distinction entre les agents alternatifs pouvant traiter le même type d'appel.

Soit $n_{i,j}^{(k)}$ le nombre d'appels de type j traités par l'agent i le jour k, où $k = 1, 2, \ldots, K_{i,j}$ et $K_{i,j}$ est le nombre total d'appels de la journée où l'agent i traite les appels de types j. Soit $\hat{m}_{i,j}^{(k)}$ le temps de service moyen d'un appel de type j traité par l'agent i le jour k, basée sur un échantillon de $n_{i,j}^{(k)}$ d'appels répondus. Notre ensemble de données contient des valeurs au jour le jour pour les deux $n_{i,j}^{(k)}$ et $\hat{m}_{i,j}^{(k)}$. Nous définissons $\hat{m}_{i,j}$ et $\hat{v}_{i,j}$ comme suit :

$$\hat{m}_{i,j} = \frac{\sum_{k=1}^{K_{i,j}} n_{i,j}^{(k)} \hat{m}_{i,j}^{(k)}}{\sum_{k=1}^{K_{i,j}} n_{i,j}^{(k)}} \tag{7.1}$$

et

$$\hat{v}_{i,j} = \frac{1}{K_{i,j} - 1} \sum_{k=1}^{K_{i,j}} n_{i,j}^{(k)} (\hat{m}_{i,j} - \hat{m}_{i,j}^{(k)})^2 \tag{7.2}$$

Ces $\hat{m}_{i,j}$ et $\hat{v}_{i,j}$ sont des estimateurs sans biais de $m_{i,j}$ et $v_{i,j}$ pour chaque agent i et type d'appel j ; voir Deslauriers (2003) pour plus de détails.

7.4.2 Modèle A1

L'analyse préliminaire en 7.3 suggère que le temps de service moyen pour un agent donné et un type d'appel n'est pas constant dans le temps ; voir la figure 7.10. Soit $M_{i,j}^{(k)}$ une variable aléatoire représentant le temps de service moyen d'un appel de type j traité par l'agent i au jour k. Ceci est ce que nous observons dans nos données. Dans le modèle A1, nous supposons que $M_{i,j}^{(k)}$ suit un processus Gaussien qui est un modèle linéaire additif à effets fixes intégrant une interception et une forme linéaire. Autrement dit, nous supposons pour chaque paire (i, j) que :

$$M_{i,j}^{(k)} = \alpha_{i,j} \cdot k + \beta_{i,j} + \epsilon_{i,j}^{(k)} \ . \tag{7.3}$$

Les coefficients $\alpha_{i,j}$ et $\beta_{i,j}$ sont des constantes réelles qui doivent être estimées à partir des données, et $\epsilon_{i,j}^{(k)}$ sont variables aléatoires normales i.i.d de moyenne 0 et variance $\sigma_{\epsilon_{i,j}}^2 / n_{i,j}^{(k)}$, où $n_{i,j}^{(k)}$ est le nombre d'appels de type j répondus le jour k par l'agent i. Autrement dit, le nombre d'appels traités dans une journée donnée est utilisé comme un poids dans notre modèle de régression. Nous estimons le modèle 7.3 utilisant la méthode des moindres carrés pondérée. Bien sûr, la modélisation de la moyenne du temps de service comme une fonction linéaire du temps ne peut avoir du sens que si l'estimation approximative s'effectue sur un intervalle de temps limité. Par exemple, la diminution de la moyenne en fonction du temps est généralement due à un effet d'apprentissage, mais cet effet va finalement être saturé et la pente de la diminution devrait se rapprocher de 0 quand le temps avance. En fait, nous allons trouver que ce modèle avec une forme linéaire est dépassé par nos deux prochains modèles, qui n'incluent pas une telle forme linéaire. En plus de la forme linéaire temporelle, nous avons également examiné les formes quadratiques et logarithmiques. Cependant, puisque les modèles avec ces nouvelles formes ne donnent pas de meilleurs résultats, nous allons seulement présenter les résultats d'une forme linéaire dans le présent document.

Nous avons constaté que l'hypothèse de normalité de la moyenne des temps de service est raisonnable dans nos données. Cela est normal puisque nos données

consistent en des moyennes quotidiennes où chaque moyenne est généralement calculée sur des dizaines de temps de service par jour. Par exemple, dans les figures 7.11 et 7.12, nous présentons les diagrammes Q-Q pour les résidus du modèle A1 pour deux agents, a_1 et a_2, qui sont ponctuels avec des bandes de confiance 95%. Les agents a_1 et a_2 ont traités de nombreux appels de type A : 8360 et 8098 appels, respectivement. Nous avons également obtenu des résultats cohérents pour les agents qui répondent à un nombre relativement faible d'appels d'un type d'appel donné, par exemple, 200-300 appels durant l'année. Pour toutes les paires agent - type d'appels, nous avons effectué le test de Lilliefors de normalité sur les résidus du modèle A1. Dans toutes ces paires, les trois premiers quartiles empiriques de la distribution des p-valeurs pour ce test sont 0.005, 0.08 et 0.3, respectivement. Dans l'ensemble, nous avons constaté qu'il n'y avait généralement pas assez de preuves statistiques pour rejeter l'hypothèse nulle que $\alpha_{i,j} = 0$. Plus précisément, les estimations empiriques des trois premiers quartiles de la distribution des valeurs de p-valeurs sont donnés par : 0.007, 0.2, et 0.5, respectivement ; en particulier, nous ne pouvions pas rejeter l'hypothèse nulle dans plus de 60% des paires agent - type d'appels. Nous avons aussi effectué des tests de Ljung-Box sur les autocorrélations des résidus pour le modèle A1, et les quartiles de la distribution empirique des p-valeurs ont été de 0.04, 0.3, et 0.6. Pour au moins 25 % des paires agent type d'appels, les autocorrélations sont statistiquement significatives au niveau de 95 %.

7.4.3 Modèle A2 : Corrélations sérielles

Capturer les dépendances entre les temps de service successifs revient à capturer les dépendances entre les moyennes des temps de service (approximativement) normaux. Les modèles à effets mixtes sont idéaux pour capturer ces dépendances avec des données plus ou moins normalement distribuées ; nous proposons maintenant un modèle de ce genre. Nous considérons un modèle linéaire à effets mixtes Gaussien pour $M_{i,j}^{(k)}$:

$$M_{i,j}^{(k)} = \beta_{i,j} + \gamma_{i,j}^{(w_k)} + \nu_{i,j}^{(k)} \tag{7.4}$$

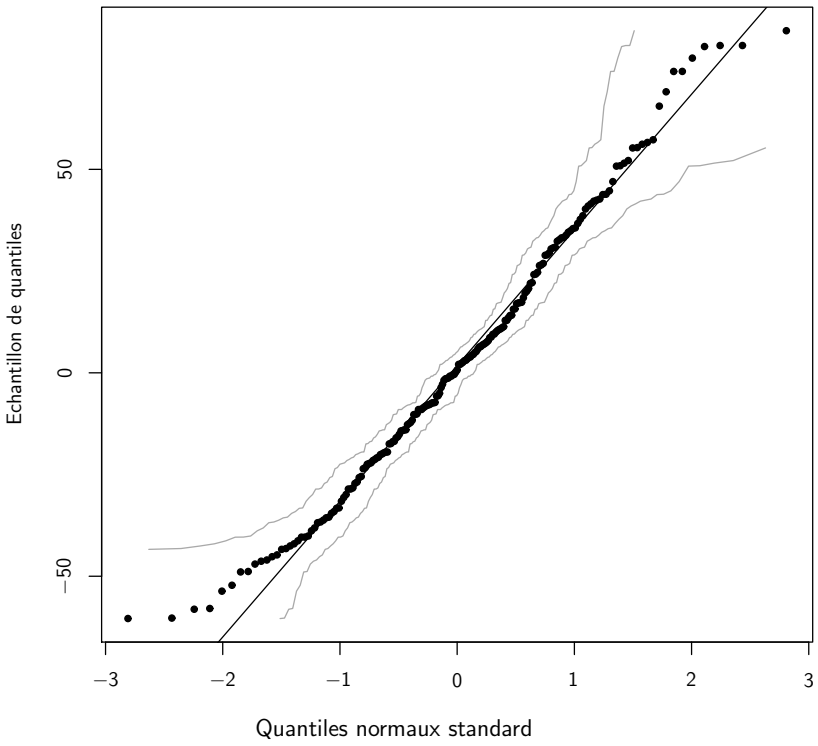

Figure 7.11 : Diagramme Q-Q des résidus du Modèle A1 pour l'agent a_1 et les bandes de confiance à 95%.

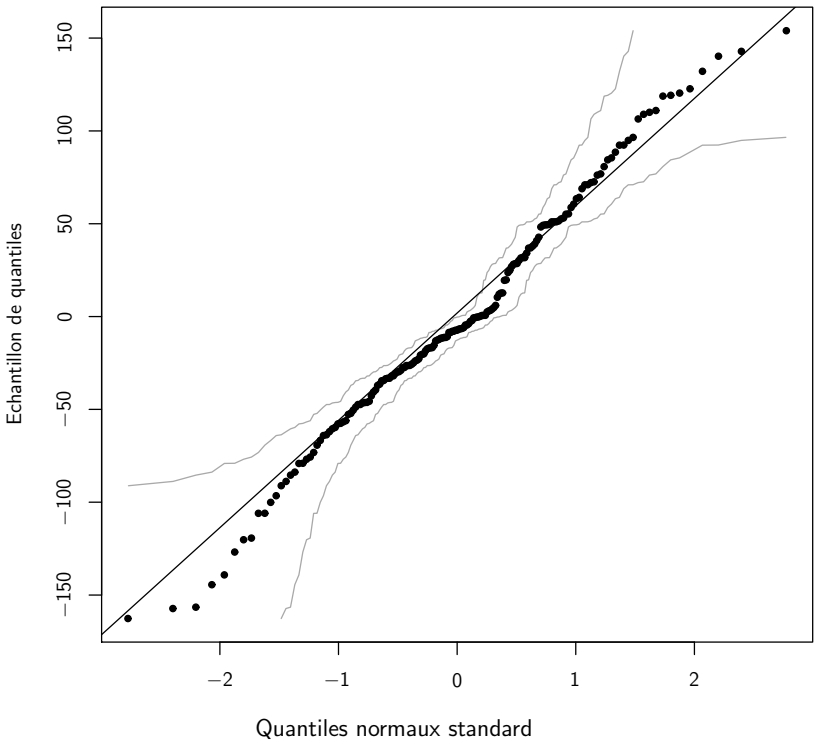

Figure 7.12 : Diagramme Q-Q des résidus du Modèle A1 pour l'agent a_2 et les bandes de confiance à 95%.

où $\gamma_{i,j}^{(w_k)}$ est un effet aléatoire spécifique pour la semaine w_k du jour k, et $\nu_{i,j}^{(k)}$ est l'erreur résiduelle normalement distribuée. Nous supposons que ces résidus $\nu_{i,j}^{(k)}$ sont des normales indépendantes de moyenne 0 et de variance $\sigma^2_{\nu_{i,j}}/n_{i,j}^{(k)}$. La variance résiduelle de $\nu_{i,j}^{(k)}$ est spécifique à chaque paire (i, j) ; ainsi, nous pouvons capturer les différences dans la variance entre les différentes paires agent/compétence. Les effets aléatoires $\gamma_{i,j}^{(w_k)}$ sont normalement distribués avec des déviations hebdomadaires que nous utilisons pour capturer des corrélations dans les temps de service moyens, pour le même agent et type d'appel, à travers les semaines successives et à travers des jours successifs de la même semaine. En raison de la nature agrégée des données disponibles, nous ne considérons pas un effet quotidien aléatoire dans (7.4), mais plutôt une par semaine, et nous n'imposons pas une structure de covariance sur les résidus $\nu_{i,j}^{(k)}$. En effet, les deux pourraient conduire à des problèmes d'identification, mais nous ne disposons pas de données pour les appels individuels au cours d'une journée donnée. Par la suite, on omet l'indice d'une variable aléatoire lorsque l'indice spécifique n'a pas d'importance. Dans les modèles à effets mixtes Gaussien, les $\gamma_{i,j}^{(w_k)}$ et $\nu_{i,j}^{(k)}$ sont censés être normalement distribués et indépendant. Ici, nous supposons que les effets aléatoires $\gamma_{i,j}^{(w_k)}$, sont des normales identiquement distribuées de valeur moyenne $E[\gamma_{i,j}^{(w_k)}]$ et de variance $Var[\gamma_{i,j}^{(w_k)}] = \sigma_{\gamma_{i,j}}$, et que $\gamma_{i,j}^{(w_k)}$ suit une structure de covariance autorégressive de premier ordre, AR(1). Ceci est,

$$\gamma_{i,j}^{(u)} = \rho_{i,j}\gamma_{i,j}^{(u-1)} + \psi_{i,j}^{(u)}\,, \tag{7.5}$$

où $\rho_{i,j}$ est un paramètre d'autocorrection et les $\psi_{i,j}^{(u)}$ sont des variables aléatoires normales i.i.d de moyenne $E[\psi_{i,j}^{(u)}] = 0$ et de variance $Var[\psi_{i,j}^{(u)}] = \sigma^2_{\psi_{i,j}^{(u)}} = \sigma^2_{\gamma_{i,j}}(1 - \rho_{i,j})$. La covariance entre $\gamma_{i,j}^{(u_1)}$ et $\gamma_{i,j}^{(u_2)}$ est donnée par

$$\mathrm{Cov}(\gamma_{i,j}^{(u_1)}, \gamma_{i,j}^{(u_2)}) = \sigma^2_{\gamma_{i,j}}\rho_{i,j}^{|u_2-u_1|}\ . \tag{7.6}$$

L'hypothèse d'une structure de covariance AR(1) pour $\gamma_{i,j}^{(w_k)}$ est à la fois utile et computationellement efficace, car elle nécessite l'estimation de deux paramètres seulement, $\sigma_{\gamma_{i,j}}$ et $\rho_{i,j}$. Ici, l'effet aléatoire hebdomadaire qui suit un processus

AR(1) remplace la forme linéaire que nous avions au modèle A1. Il permet une situation où, pour un agent donné, par exemple, la moyenne diminue pendant une certaine période de temps en raison de l'apprentissage, puis reste stable, puis augmente plus tard parce que l'agent perd intérêt ou a d'autres problèmes, etc. Ce processus AR(1) est très simple et pourtant suffisamment flexible pour modéliser ces variations à mi-parcours dans la moyenne.

Nous avons aussi essayé le modèle $A2$ avec une forme linéaire comme dans $A1$ en plus du terme AR(1), et avons constaté que la version sans la forme linéaire fournit un meilleur ajustement aux données sur l'échantillon. Pour cette raison, nous avons omis la forme linéaire. Dans le tableau 7.1, nous présentons des estimations ponctuelles pour les différents paramètres du modèle $A2$, basé de nos données, pour 3 combinaisons agent/type d'appel. Les p-valeurs dans le tableau sont calculées automatiquement dans SAS® comme suit : En supposant la normalité des effets aléatoires et des résidus, nous pouvons construire une statistique en fonction des effets fixes (β) et des effets aléatoires (γ) qui a approximativement une loi t pour laquelle nous pouvons estimer les degrés de liberté. Basés sur cette statistique, nous pouvons faire une inférence pour savoir si les effets aléatoires et fixes (la forme linéaire) sont égaux à 0. L'effet aléatoire hebdomadaire et le paramètre d'autocorrélation sont généralement jugés statistiquement significatifs. Lors du test de forme linéaire, les quartiles de la distribution empirique des p-valeurs étaient 0.06, 0.3, et 0.6. La forme est statistiquement significative avec un niveau de 95 % pour 125 paires sur 550. Pour le modèle sans la forme linéaire, l'autocorrélation est statistiquement significative au niveau 95% pour 246 paires sur 550, et les quartiles de la distribution des p-valeurs sont 0, 0.002 et 0.3.

7.4.4 Modèle A3 : Corrélations sérielle et croisée

Les dépendances dans les séries de temporelle des temps de service peuvent être dues à des facteurs liés aux agents eux-mêmes, tel que le stress, la fatigue, la démotivation, etc. Ces effets à court terme peuvent influencer les performances de l'agent pendant une période de temps donnée et provoquer des dépendances entre

(Agent,type d'appel)	Catégorie	Valeur	Erreur std.	p-valeur
(i_0, j_0)	$\sigma^2_{\gamma_{i_0,j_0}}$	1270	1004	0.1035
	ρ_{i_0,j_0}	0.705	0.269	0.00870
	$\sigma^2_{\epsilon_{i_0,j_0}}$	146000	20800	$< .0001$
	β_{i_0,j_0}	602	45.7	$< .0001$
	α_{i_0,j_0}	-0.634	0.204	0.00250
(i_1, j_1)	$\sigma^2_{\gamma_{i_1,j_1}}$	608	356	0.0439
	ρ_{i_1,j_1}	0.870	0.0846	$< .0001$
	$\sigma^2_{\epsilon_{i_1,j_1}}$	93867	10300	$< .0001$
	β_{i_1,j_1}	295	21.9	$< .0001$
	α_{i_1,j_1}	0.0885	0.101	0.383
(i_2, j_2)	$\sigma^2_{\gamma_{i_2,j_2}}$	1320	684	0.0267
	ρ_{i_2,j_2}	0.652	0.244	0.00760
	$\sigma^2_{\epsilon_{i_2,j_2}}$	51000	8030	$< .0001$
	β_{i_2,j_2}	283	25.4	$< .0001$
	α_{i_2,j_2}	0.243	0.156	0.124

Tableau 7.1 : Résultats pour le Modele $A2$ pour 3 différentes combinaisons agent/type d'appel. Les estimations ponctuelles des coefficients du modèle sont montrées avec les erreurs standard et p-valeurs pour des significations statistiques des t-tests.

les temps de service de tous les appels traités par ce même agent. La considération des modèles avec des corrélations croisées est donc importante pour capturer des effets similaires.

Dans le modèle A3, nous modélisons conjointement les temps de service des différents types d'appels traités par le même agent. Nous considérons un modèle à effets mixtes pour la moyenne des temps de service (tout comme dans le modèle A2) où nous fusionnons les types d'appels alternatifs ensemble et avons le même effet aléatoire hebdomadaire commun à tous les types traités par le même agent. Cela donne :

$$M^{(k)}_{i,j} = \beta_{i,j} + \gamma^{(w_k)}_i + \nu^{(k)}_{i,j} \tag{7.7}$$

Le terme constant $\beta_{i,j}$ est spécifique au type d'appel j traité par l'agent i. Nous continuons à assumer une structure de covariance AR(1) pour $\gamma^{(w_k)}_i$. Notons que

$\gamma_i^{(w_k)}$ dépendra de l'agent i et la semaine w_k, mais pas du type d'appel j. Nous continuons également à supposer que les résidus sont des normals i.i.d de moyenne 0 et de variance $\sigma_{\nu_{i,j}}/n_{i,j}^{(n)}$. L'effet aléatoire $\gamma_i^{(w_k)}$, qui est commun à tous les types d'appels traités par l'agent i, exploite à la fois la corrélation sérielle entre les semaines successives, et les corrélations croisées entre les différents types d'appels. La variance résiduelle $\nu_{i,j}^{(k)}$ est spécifique à chaque paire $(i,\, j)$; ainsi nous capturons les différences de variance entre les différentes paires (agent,compétence).

Pour illustrer, le tableau 7.2, donne les paramètres estimés du modèle $A3$ pour l'agent i_0 considéré dans le tableau 7.1. Ici, la p-valeur pour l'effet aléatoire hebdomadaire est 0.0858. D'autres p-valeurs sont assez petites. Nous avons également testé le modèle $A3$ avec une forme linéaire, pour notre cohorte de 200 agents, et les quartiles de la distribution des p-valeurs pour le test de forme linéaire étaient de 0.04, 0.3, et 0.6. Autrement dit, pour la plupart des agents la forme linéaire est non significative. Dans les tests de la qualité de l'ajustement hors échantillon et des prédictions, basés sur le modèle $A3$ avec et sans la forme linéaire, la version sans la forme linéaire s'ajuste mieux aux données. Par conséquent, nous omettons cette forme linéaire des considérations en §7.5 et 7.6. Pour le modèle sans la forme linéaire, le paramètre d'autocorrélation est généralement jugé statistiquement significatif : les quartiles de la distribution des p-valeurs étaient (approximativement) 0, 0.005 et 0.2.

7.5 Qualité de l'ajustement des modèles

Dans cette section, nous évaluons la qualité de l'ajustement aux données de nos modèles candidats.

7.5.1 Modèle des résidus

Nous commençons par l'analyse des résidus de chaque modèle, où les résidus du modèle sont définis comme étant égaux à la différence entre la moyenne des temps de service quotidiens observée et les valeurs ajustées correspondantes. Dans

Categorie	Valeur	Erreur stand.	p-valeur
$\sigma^2_{\gamma_{i_0}}$	1240	901	0.0858
$\rho_{\gamma_{i_0}}$	0.687	0.270	0.0108
$\sigma^2_{\epsilon_{i_0,1}}$	146000	47200	0.001
$\sigma^2_{\epsilon_{i_0,2}}$	149000	32300	< 0.0001
$\sigma^2_{\epsilon_{i_0,3}}$	145000	19800	< 0.0001
$\beta_{(i_0,1)}$	562.1	107	< 0.0001
$\beta_{(i_0,2)}$	454	43.2	< 0.0001
$\beta_{(i_0,3)}$	624	42.7	< 0.0001
$\alpha_{(i_0,1)}$	-0.628	0.686	0.361
$\alpha_{(i_0,2)}$	-0.371	0.193	0.0564
$\alpha_{(i_0,3)}$	-0.727	0.192	0.0002

Tableau 7.2 : Résultats pour le Modele $A3$ pour l'agent i_0, présenté dans le tableau 7.1, répondant à 3 different types d'appels, numérotés de 1 à 3. Les estimations ponctuelles des coefficients du modèle sont montrés avec les erreurs standard et p-valeurs indiquant ce qui est statistiquement significatif.

le tableau 7.3, nous présentons un résumé des statistiques pour le carré des résidus pour notre cohorte C des agents ; voir 7.3. Le tableau 7.3 montre que les modèles $A2$ et $A3$ s'ajustent mieux aux données que le modèle $A1$, $B1$ et $B2$, et que le modèle $A2$ donne un ajustement un peu meilleur que le modèle $A3$. Les Modèles $B1$ et $B2$ sont dernière, et le modèle $B1$ donne clairement le plus mauvais ajustement.

Nous calculons également les estimations du RMSE pour la cohorte C sous les différents modèles. Dans la figure 7.13, nous présentons des boîtes de moustaches pour les RMSEs à travers tous les modèles. La Figure 7.13 montre que les modèles $A2$ et $A3$ s'ajustent mieux aux données que le reste des modèles. Dans la figure 7.14, nous traçons les ECDFs des RMSEs pour tous les modèles. Une fois de plus, la figure 7.14 montre que les modèles $A2$ et $A3$ fournissent les meilleurs ajustements aux données que les autres modèles. Pour les RMSEs, nous avons effectué des tests t par paire, avec un niveau de confiance de 95 %, pour toutes les paires du modèle et avons constaté que les différences dans les RMSE étaient toutes significativement différentes de 0 (les p-valeurs correspondants de tous les tests ont été très proches

de 0).

7.6 Prédictions de la moyenne des temps de service

Nous comparons maintenant les modèles statistiques de 7.4 basé sur leur performance de prédictions hors échantillon, pour notre cohorte de C agents. Pour chaque agent et type d'appel, chaque modèle, et chaque jour i, nous avons estimé le modèle basé seulement sur toutes les observations jusqu'au jour $i - \delta$ (la période d'apprentissage), où δ est le temps (nombre de jours) de prédiction en avance choisie ou *"lead time"*, et à partir de là nous avons calculé une prédiction $\hat{m}_i$ de la moyenne du temps de service m_i pour la journée i. Nous avons considéré seulement les jours i pour lesquels $i - \delta \geq 60$. Chaque $\hat{m}_i$ est une prédiction hors échantillon (basée uniquement sur des informations passées). Nous considérons trois δ différents, à savoir 2 semaines, 1 semaine et 1 jour, pour imiter les défis de la vie réelle auxquels sont confrontés les gestionnaires de centres d'appels. Nous déterminons la période d'apprentissage en avance de manière à préserver la condition $i - \delta \geq 60$. Nous réestimons tous les paramètres du modèle après chaque prédiction. Nous utilisons la procédure mixte SAS® pour calculer les estimations du maximum de vraisemblance des paramètres pour le modèle $A2$ et modèle $A3$, et générer les prédictions correspondantes.

Dans les tableaux 7.4 et 7.5, nous rapportons les résultats agrégés pour les prévisions hors échantillon sur tous les 200 agents, types d'appel, et jours. Dans le tableau 7.4, nous incluons des estimations de la moyenne, la médiane, et le premier et troisième quartile des MAPEs et RMSEs obtenus à travers tous les agents. Rappelons que chaque MAPE et RMSE est sur tous les types d'appel et jours, pour chaque agent. Nous soulignons en gras le RMSE et MAPE minimum dans chaque rangée. Il est clair que le modèle $A3$ est supérieur. Il surpasse clairement nos modèles de référence, couramment utilisés dans la pratique, en particulier avec des prévisions en avance très courtes (un jour). Nous discutons maintenant brièvement des résultats pour des prévisions en avance de 2 semaines et 1 jour, respectivement.

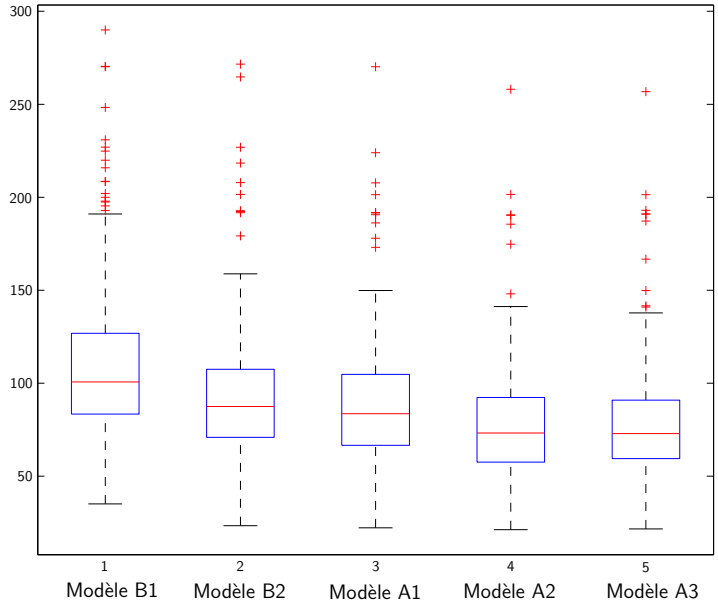

Figure 7.13 : Boîte de moustaches du RMSE du modèle des résidus lors de l'ajustement de tous les modèles aux données de la cohorte de C agents.

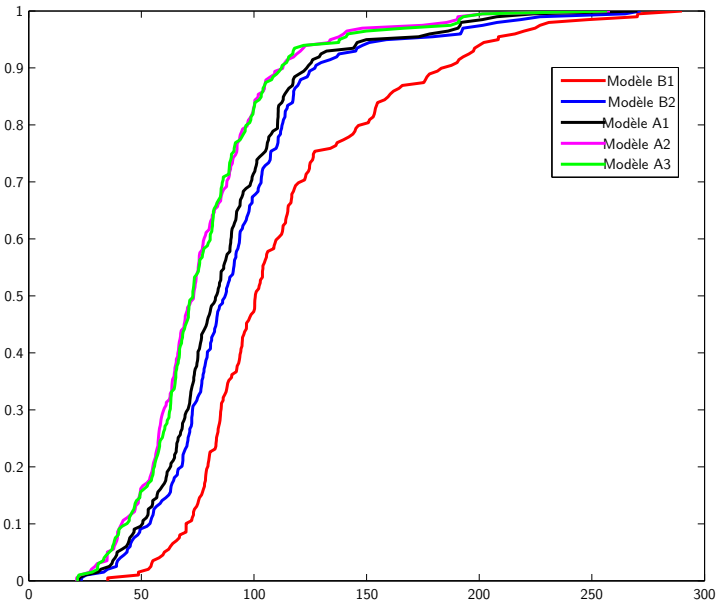

Figure 7.14 : ECDF pour le RMSE des modèles de résidus lors de l'ajustement aux données à la cohorte de C agents.

Statistique	Modèle $B1$	Modèle $B2$	Modèle $A1$	Modèle $A2$	Modèle $A3$
Moyenne	15,243	10,131	9,214	7,272	7,428
Mediane	4,238	2,296	2,099	1,574	1,624
Premier quartile	894	454	415	308	322
Troisième quartile	13,396	7,844	7,217	5,581	5,690

Tableau 7.3 : Résumé statistique du carré des résidus avec chaque modèle, à travers la cohorte de C agents.

7.6.1 Prédictions de deux semaines en avant

Pour cette longue période, les modèles $A2$ et $A3$ donnent à peu près les mêmes performances. Le modèle $B2$ est aussi bien compétitif, et donne des erreurs de prédiction plus petites que le modèle $A1$, tant pour le MAPE et la RMSE. Le MAPE moyen pour le modèle $A3$ est d'environ 12% inférieur à celui de $A1$ et 3% inférieur à celui de $B2$. D'autre part, le modèle $B1$ accuse un net retard, avec un MAPE 26% plus grand que pour le modèle $A3$. Des résultats similaires sont obtenus quand nous comparons le RMSE moyen, qui est d'environ 18 % plus grandes pour le modèle $A1$ que pour $A3$ et 3% plus grande pour $B2$ que pour $A3$. Nous avons effectué des tests t par paire, pour toutes les paires possibles, pour tester si les différences dans la RMSE et MAPE sont significativement différentes de 0. Les p-valeurs de ces tests étaient tous très petits (la différence est clairement significative, avec $p < 0.0001$), sauf dans deux cas : Lorsque nous comparons les MAPEs pour les modèles $B1$ et $A1$, nous obtenons $p = 0.48$ et en comparant le MAPE pour le modèle $A2$ et $A3$, on obtient $p = 0.12$.

7.6.2 Prédictions d'une journée en avant

Avec une prévision en avance de 1 jour, l'avantage d'exploiter les corrélations entre les moyennes hebdomadaires augmente. Les modèles $A1$, $A2$, et $A3$ donne des prévisions plus précises que les deux modèles de référence, avec le modèle $A3$ qui prend la tête. Par exemple, le MAPE moyen est d'environ 6% plus petit pour le modèle $A3$ que pour $B2$, et environ 5% plus petit pour $A2$ que pour $A1$. Le RMSE

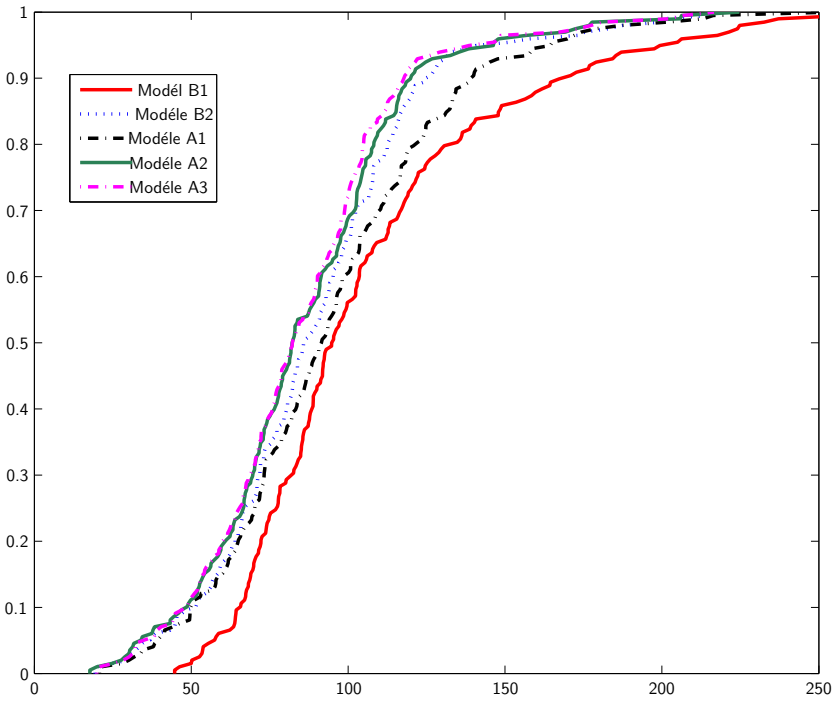

Figure 7.15 : ECDF pour le RMSE pour les prévisions d'une journée à l'avance, à travers tous les agents de la cohorte C.

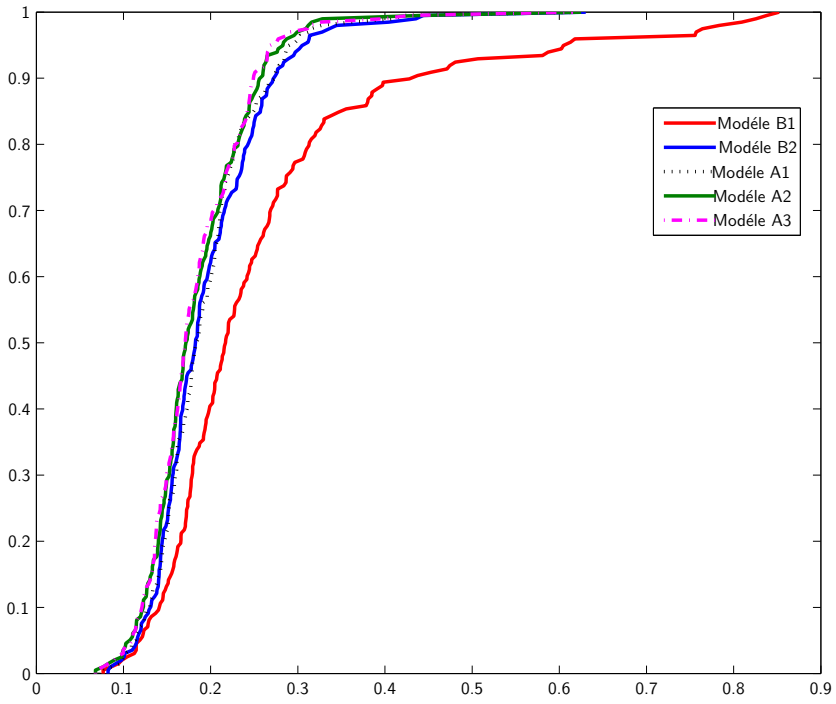

Figure 7.16 : ECDF pour le MAPE pour les prévisions d'une journée à l'avance, à travers tous les agents de la cohorte C.

Prévision de deux semaines à l'avance.

	Modèle $B1$		Modèle $B2$		Modèle $A1$		Modèle $A2$		Modèle $A3$	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Moyenne	107	26.2	91.5	20.1	109	22.0	89.6	19.6	**88.7**	**19.4**
Mediane	95.2	21.7	88.1	18.6	97.4	20.7	86.7	18.0	**86.0**	**17.9**
Premier quartile	77.3	17.3	70.0	15.3	74.7	16.5	69.1	15.0	**68.2**	**14.9**
Troisième quartile	122.0	28.9	109	23.3	136	25.7	108	22.8	**106**	**22.8**

Prévision d'une semaine à l'avance

	Modèle $B1$		Modèle $B2$		Modèle $A1$		Modèle $A2$		Modèle $A3$	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Moyenne	107	26.2	90.6	19.9	100.0	20.4	88.3	19.2	**87.2**	**19.0**
Mediane	94.7	21.8	87.6	18.5	96.1	19.4	85.0	17.8	**84.2**	**17.6**
Premier quartile	77.3	17.4	68.5	15.3	73.5	15.9	67.9	15.0	**67.4**	**14.7**
Troisième quartile	122	28.8	109	23.3	124	23.5	107	**22.3**	**105**	22.7

Prévision d'une journée à l'avance

	Modèle $B1$		Modèle $B2$		Modèle $A1$		Modèle $A2$		Modèle $A3$	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Moyenne	107	26.2	89.9	19.6	95.2	19.4	86.6	18.5	**85.4**	**18.4**
Mediane	95.1	21.7	86.1	18.3	91.6	18.4	82.3	17.2	**82.3**	**17.1**
Premier quartile	77.6	17.4	67.6	15.2	70.4	15.3	66.8	14.6	**65.9**	**14.2**
Troisième quartile	122	28.7	108	23.2	117	22.0	104	**21.6**	**102**	21.7

Tableau 7.4 : Précision des prédictions pour les Modèles $A1$, $A2$, et $A3$, moyenne à travers la cohorte de C agents.

moyen est d'environ 10 % plus petits pour le modèle $A3$ que pour $A1$, et environ 9 % plus petite pour $A2$ que pour $A1$. Le modèle $B1$ est clairement dépassé par tous les autres modèles. Dans des t-tests de paires, tous les p-valeurs étaient très petits ($p < 0.0001$).

7.6.3 Proportion de victoires pour chaque modèle

Le tableau 7.5 compare les modèles à partir d'un point de vue différent : il rapporte les proportions d'agents (dans notre cohorte C) où chaque modèle donne la mesure de performance la plus petite, dans tous les modèles. Par exemple, la première ligne du tableau 7.5 indique que, à travers les 200 agents considérés, le plus petit MAPE a été réalisé par le modèle $A1$ pour 20.2% des agents, par le

Prévision de deux semaines à l'avance.

	Modèle $B1$	Modèle $B2$	Modèle $A1$	Modèle $A2$	Modèle $A3$
MAPE	0.202	0.227	0.202	0.157	**0.253**
RMSE	0.167	0.182	0.106	0.212	**0.364**

Prévision d'une semaine à l'avance

	Modèle $B1$	Modèle $B2$	Modèle $A1$	Modèle $A2$	Modèle $A3$
MAPE	0.172	0.187	0.227	0.202	**0.278**
RMSE	0.141	0.152	0.131	0.232	**0.389**

Prévisions d'une journée à l'avance

	Modèl $B1$	Modèle $B2$	Modèle $A1$	Modèle $A2$	Modèle $A3$
MAPE	0.141	0.146	0.207	0.232	**0.338**
RMSE	0.116	0.116	0.136	0.278	**0.424**

Tableau 7.5 : Les proportions où un modèle donné est gagnant, c.-à-d., donne les plus petites mesures de performances, à travers la cohorte de C agents.

modèle $A2$ 15.7 % des agents, et par le modèle $A3$ pour 25.3% des agents. Avec un temps de prévision en avant de deux semaines, le modèle $B2$ est compétitif, mais il est encore dépassé par le modèle $A3$. Dans le tableau 7.5, le modèle $A3$ est généralement plus performant que tous les autres modèles. Cela est particulièrement vrai avec un temps de prévision en avant très court. Par exemple, avec un délai de prévision d'un jour, le modèle $A3$ donne le plus petit RMSE pour 42.4% des agents, par rapport à 11.6% pour $B2$.

Dans les figures 7.15 et 7.16, nous traçons les fonctions de répartition empiriques pour le RMSE et la MAPE, respectivement, pour tous les modèles, avec un délai de prévision en avant d'un jour. Ces figures illustrent l'amélioration de la précision des prévisions, qui est résumée dans le tableau 7.4.

7.7 Simulation

Dans les sections précédentes, nous avons illustré l'amélioration de la qualité de l'ajustement aux données qui résulte de la considération de modèles de temps de service plus réalistes. Nous allons maintenant discuter les résultats des études de

simulation qui permettent d'évaluer l'impact de considérer les différents modèles de temps de service sur les mesures de performances dans un centre d'appels. Nous avons considéré le temps d'attente moyen (AWT) des appels et le niveau de service (SL). Nous avons choisi de simuler seulement les vendredis, car pour ces journées l'estimation du γ de la semaine des modèles A2 et A3 devrait avoir moins de bruit dû à la quantité de données observées pour son estimation.

Nous allons d'abord décrire comment les paramètres des centres d'appels étudiés sont estimés. Dans la section 7.7.1, nous allons présenter comment sont simulés les processus d'arrivée, les temps de patience et les durées de service avec les différents modèles considérés qui sont B1, B2, A2 et A3.

7.7.1 Estimations des paramètres

7.7.1.1 Les processus d'arrivée

Le processus d'arrivée pour chaque type d'appel est Poisson par morceaux, avec un taux d'arrivée aléatoire gamma dans chaque intervalle de temps de 15 minutes, et une copule normale pour modéliser la dépendance entre ces taux. Ce modèle est bien expliqué dans Oreshkin et al. (2016), où il y est également montré que ces processus d'arrivées s'ajustent bien aux données d'arrivées pour ce centre d'appel HQ. Nous avons utilisé les données recueillies au centre d'appels pour estimer les paramètres du processus des différents types d'appels utilisés dans nos simulations.

7.7.1.2 Les temps de patience

Les temps de patience pour chaque type d'appel j du centre d'appels HQ sont exponentiels de moyenne ν_j^{-1} estimée également sur des données réelles recueillies au centre d'appels HQ.

7.7.1.3 Temps des temps de service

Soit $\mathcal{F}_{i,j}^{(k)}$ toutes les informations dans les données jusqu'au jour k sur les paires (i, j).

Modèle B1 et B2. Pour simuler les temps de service des agents à la journée $(k+1)$ avec les modèles B1 et B2, nous estimons la moyenne et la variance des distributions log-normales pour les temps de service en utilisant $\mathcal{F}_{i,j}^{(k)}$ à l'aide des formules 7.1 et 7.2 de la section 7.4.1 pour prédire ces paramètres à la journée $(k+1)$.

Modèle A2. Notons que $\mathcal{F}_{i,j}^{(k)}$ ne contient pas les valeurs de $\gamma_{i,j}^{(w_k)}$, puisque ceux-ci ne peuvent pas être observées. Avec le modèle A2, nous savons que la moyenne des temps de service au jour $k+1$ est donné par :

$$M_{i,j}^{(k+1)} = \beta_{i,j} + \gamma_{i,j}^{w_{k+1}} + \nu_{i,j}^{(k+1)},$$

et si nous désignons par $\hat{M}_{i,j}^{(k+1)} = \hat{\beta}_{i,j} + \hat{\gamma}_{i,j}^{w_{k+1}}$ la valeur prédite de $M_{i,j}^{(k+1)}$ conditionnelle à $\mathcal{F}_{i,j}^{(k)}$, où $\hat{\beta}_{i,j}$ est la valeur estimée de $\beta_{i,j}$ et $\hat{\gamma}_{i,j}^{w_k}$ la valeur de $\gamma_{i,j}^{w_k}$ prédite au jour k.

Pour simuler les temps de service $S_{i,j}^{(k+1,l)}$ d'un agent i au jour $k+1$ pour le type d'appel j, conditionnelle à $\mathcal{F}_{i,j}^{(k)}$, nous simulons en premier $\mu_{i,j}^{(k+1)}$ qui est normalement distribué de moyenne $\hat{M}_{i,j}^{(k+1)}$ et de variance $v_{i,j}^{(k)}$. Notons par $m_{i,j}^{(k+1)}$ cette valeur simulée. Ensuite, nous générons les valeurs $S_{i,j}^{(k+1,l)}$ qui représentent les temps de service de l'agent i pour le type d'appel j par la distribution log-normale d'espérance $m_{i,j}^{(k+1)}$ et de variance $\sigma_{i,j}^2 = n_{i,j}^{(k)}\sigma_{\nu_{i,j}}^2$. Notons que $v_{i,j}^{(k)}$ désigne une estimation de $E[(\gamma_{i,j}^{(w_k)} - \hat{\gamma}_{i,j}^{(w_k)})^2]$ qui représente la variance sur l'estimation du $\hat{\gamma}_{i,j}^{(w_k)}$.

Pour l'estimation de $v_{i,j}^{(k)}$, nous pouvons utiliser deux méthodes différentes que nous appelons **Méthode 1** et **Méthode 2**. Pour la première méthode, nous utilisons une formule mathématique et pour la deuxième nous développons une heuristique.

Méthode 1 :

Notons que :

$$E[(M_{i,j}^{(k)} - \hat{M}_{i,j}^{(k)})^2] = E[((\beta_{i,j}^{(k)} - \hat{\beta}_{i,j}^{(k)}) + (\gamma_{i,j}^{(w_k)} - \hat{\gamma}_{i,j}^{(w_k)}) + \nu_{i,j}^{(k)})^2].$$

Nous pouvons calculer une estimation de $E[(M_{i,j}^{(k)} - \hat{M}_{i,j}^{(k)})^2]$ pour la journée k à partir de données : Ceci est l'espérance du carré de la différence entre le temps de service moyen observé et sa valeur prédite au jour k, c'est-à-dire le MSE. Désignons par $e_{i,j}^{(k)}$ cette estimation du MSE. Ainsi, sans tenir compte de l'erreur dans l'estimation de $\beta_{i,j}$, une estimation de $E[(\gamma_{i,j}^{(w_k)} - \hat{\gamma_{i,j}}^{(w_k)})^2]$ est donnée par :

$$v_{i,j}^{(k)} = e_{i,j}^{(k)} - \sigma_{\nu_{i,j}}^2/N,$$

où N est la variable aléatoire du nombre d'appels de j répondus par l'agent i à la journée k. Pour chaque paire (agent i, type d'appel j), nous pouvons estimer $e_{i,j}^{(k)}$ de deux manières différentes. Dans premier cas, nous estimons $e_{i,j}^{(k)}$ en calculant le MSE observé sur tous les jours en utilisant la moyenne du temps de service observée et la moyenne prédite. Dans le deuxième cas, nous estimons $e_{i,j}^{(k)}$ en calculant le MSE en utilisant les moyennes observées et prédites pour les vendredis seulement. Dans les deux cas, $e_{i,j}^{(k)}$ ne dépend pas de k. Ainsi pour l'estimation du v_k à la journée k, nous utilisons la formule suivante

$$v_{i,j}^{(k)} = e_{i,j}^{(k)} - \sigma_\nu^2 \times E[1/N].$$

Mais l'utilisation de cette méthode d'estimation n'a pas fonctionné pour tous les couples (i, j) car elle donne des valeurs aberrantes avec certains agents. Ce qui nous fait dire qu'il y a eu beaucoup de bruit sur l'estimation de $\sigma_{\nu_{i,j}}^2$.

Méthode 2 :

Nous avons développé une heuristique qui pour estimer $v_{i,j}^{(k)}$ pour un jour de vendredi utilise la distribution des erreurs sur l'estimation du $\gamma_{i,j}^{w_k}$, et pour toute autre journée

différente du vendredi exploite la structure du processus AR(1) pour estimer $v_{i,j}^{(k)}$. Pour simplifier l'écrire, nous utiliserons par la suite γ à la place de $\gamma_{i,j}^{w_k}$ et v_k au lieu de $v_{i,j}^{(k)}$.

Nous savons que l'erreur sur l'estimation du γ de la semaine w_k diminue au fur et à mesure que l'on avance dans la semaine, car d'autant plus que l'on avance dans la semaine d'autant plus nous disposons de plus de données. La disponibilité de beaucoup plus de données permet d'avoir une bonne estimation du γ. Nous voulons dire par là qu'il y a moins d'erreurs (en moyenne) dans l'estimation du γ de la semaine le jeudi soir que le lundi matin. En plus, à la fin de la journée du jeudi, nous avons observé plus de 80% des données nécessaire pour l'estimation du γ de la semaine w_k donc nous pouvons considérer que l'erreur sur l'estimation du γ est négligeable et considérer que $\hat{\gamma} \approx \gamma$. Avec cette supposition, nous pouvons ainsi facilement déterminer la distribution des erreurs sur l'estimation du γ de la semaine à la fin de la journée du jeudi. Ainsi, nous pouvons utiliser cette distribution des erreurs pour calculer la variance empirique de v_k pour la journée du vendredi.

À la fin de la journée du vendredi, nous pouvons à nouveau ré-estimer $\hat{\gamma}$ qui devrait correspondre au vrai γ de la semaine, donc plus précis que celui estimé au jeudi soir. Cette dernière valeur de $\hat{\gamma}$ peut être utilisée avec le processus AR(1) pour la prédiction du v_k pour les journées du lundi au jeudi de la semaine prochaine. Ainsi pour toute journée différente du vendredi, nous exploitons la structure du processus AR(1) pour estimer v_k. Ainsi nous obtenons que $v_k = \rho^2.v_{k-1} + \sigma_\psi^2$. pour les vendredis. Nous avons utilisé cette méthode pour estimer v_k dans nos simulations.

Modèle A3. Le modèle A3 est similaire au modèle A2, sauf que, dans le modèle A3, on fusionne les types d'appels alternatifs d'un agent ensemble et avons le même effet aléatoire hebdomadaire commun à tous les types traités par le même agent. Nous remplaçons $\gamma_{i,j}$ par γ_i et $\mathcal{F}_{i,j}^{(k)}$ par $\mathcal{F}_i^{(k)}$ dans les formules du modèle A2.

Les tableaux 7.6 et 7.7 montrent les valeurs estimées de v_k avec la Méthode 1 et avec la Méthode 2 respectivement, pour certaines paires (i, j) au vendredi de

la 45 ième semaine de nos données. Dans chacun des tableaux, nous rapportons, les valeurs de v_k estimée en utilisant seulement les données des vendredis, mais aussi les v_k estimés en utilisant toutes les données. Nous rapportons aussi dans ces tableaux les valeurs estimées de e_k, σ_ν^2, et E[1/N] qui sont utilisés pour estimer v_k. Comme espéré, nous observons dans les deux cas que les valeurs du MSE, e_k, sont en général plus petites en utilisant seulement les données des vendredis qu'en utilisant les données de toutes les journées. Cependant nous constatons avec la Méthode 1, les v_k estimés sont souvent négatifs que ce soit avec les données des vendredis que ce soit avec les données sur toutes les journées. Ainsi nous pensons que si la méthode d'estimation de e_k est correcte alors il y a beaucoup de biais dans l'estimation σ_ν^2.

Agent	skill	Tous les Vendredis				Tous les jours			
		e_k	σ_ν^2	$E[1/n]$	v_k	e_k	σ_ν^2	$E[1/N]$	v_k
1	F	5941	210088	0.062	-7236	16390	210088	0.064	2806
2	F	6450	172225	0.072	-6017	10092	172225	0.071	-2163
3	F	1666	113044	0.032	-2054	3309	113044	0.039	-1175
4	F	4095	115399	0.043	-899	4708	115399	0.045	-503
5	F	17877	222747	0.055	5467	15601	222747	0.058	2516
6	F	5534	131373	0.050	-1107	8195	131373	0.042	2664
7	F	5593	83064	0.050	1392	6157	83064	0.049	2005
8	F	3495	269650	0.034	-5879	10946	269650	0.040	101
9	F	2596	90411	0.052	-2150	6260	90411	0.056	1149
10	F	3609	186808	0.46	-4985	4888	186808	0.042	-3061
11	F	15102	129976	0.053	-8204	8680	129976	0.055	1456
12	F	8956	130946	0.056	1445	7729	130946	0.059	345
11	E	15102	129976	0.053	8204	8680	129976	0.055	1456
12	E	9356	128232	0.056	1445	7729	128232	0.059	745

Tableau 7.6 : Les valeurs de v_k estimé avec la méthode 1, e_k, σ_ν^2 , et E[1/N] pour certains agents au vendredi de la semaine 45 pour le modèle A3.

7.7.2 Impact des différents modèles de temps de service sur les performances

Notre objectif dans cette section est de quantifier les différences dans les performances moyennes du système à travers nos modèles alternatifs de temps de service.

Agent	skill	M_k	$\hat{M}_k$				σ_k^2				v_k	
			B1	B2	A2	A3	B1	B2	A2	A3	A2	A3
1	F	549	415	563	562	562	1895421	261328	209593	210088	80	76
2	F	649	415	500	500	501	1895421	138991	134835	172225	83	94
3	F	280	415	321	299	284	1895421	189993	107504	113044	38	81
4	F	523	415	367	386	395	1895421	145158	128098	115399	20	18
5	F	538	415	443	442	445	1895421	262364	220645	222747	60	67
6	F	427	415	461	422	430	1895421	244324	127759	131373	81	82
7	F	342	415	369	354	330	1895421	131124	73490	83064	281	168
8	F	446	415	480	447	449	1895421	296428	212969	269650	164	10
9	F	397	415	419	408	414	1895421	134825	105224	90411	74	152
10	F	387	415	424	424	424	1895421	200555	202556	186808	40	12
11	F	385	415	362	388	401	1895421	149662	126127	129976	92	94
12	F	453	415	409	487	499	1895421	105126	127125	130946	48	64
11	E	417	412	378	428	415	331230	157662	125483	129976	158	100
12	E	444	412	456	456	451	331230	158621	124253	128232	59	56

Tableau 7.7 : Les valeurs de v_k estimé avec la méthode 2, e_k, σ_ν^2 , et E[1/N] pour certains agents au vendredi de la semaine 45 pour le modèle A3 .

Ainsi, nous montrons que la modélisation efficace des temps de service est également importante d'un point de vue opérationnel. Pour ce faire, nous considérons un modèle d'une petite partie du véritable centre d'appel HQ à partir de laquelle nos données ont été prises.

Nous considérons un modèle avec deux types d'appels et deux groupes d'agents, qui constitue un modèle N pour le routage des appels. Les deux types d'appels sont pour le même type de service (liés à la facturation), mais l'un est en Français (F) et l'autre en Anglais (E). Le premier groupe (F, avec 10 agents, numérotés de 1 à 10) ne peut traiter que le premier type d'appel et le deuxième groupe (EF, avec 2 agents, numérotée 11 et 12) peut traiter les deux types. Ainsi, les agents de F ne répondent qu'aux appels en Français et les agents de EF sont bilingues. Ces 12 agents sont ceux qui ont travaillé sur chacune des journées de la semaine numéro 45 dans notre ensemble de données et traité seulement ces deux types d'appels.

Le centre d'appels est ouvert de 8h à 18h. La journée est divisée en 40 périodes de 15 minutes. Les taux d'arrivée réels des deux types d'appels ont été revus à la baisse pour tenir compte de notre petit nombre d'agents. Les tableaux 7.15 et

7.16 donnent le taux et la forme de la distribution gamma du processus d'arrivée, respectivement, pour le type d'appel F et E. Les tableaux 7.17, 7.18, 7.19, et 7.20 donnent respectivement la matrice de corrélation entre les taux pour les types F et le type E. Le taux d'abandon est de 1.87 appel par heure pour le type d'appel F et le taux d'abandon est de 1.57 appel par heure pour le type d'appel E. Nous utilisons une politique de routage qui fonctionne comme suit. Les appels du même type sont de premier arrivé, premier servi. Pour un type d'appel F, le routeur va d'abord essayer de l'affecter à un agent F libre. S'il n'y a pas d'agent libre, alors le routeur va essayer de l'affecter à un agent EF libre. Les agents EF donnent la même priorité aux deux types d'appels. Si un agent de ce groupe devient libre et il y'a des appels dans les deux files d'attente, la priorité est donnée à l'appel qui a attendu le plus longtemps. Le vecteur de staffing pour les 40 périodes pour les agents F est donné dans le tableau 7.14. Le vecteur de staffing pour les agents EF est 2 dans toutes les périodes.

La journée cible que nous voulons simuler est le vendredi de la semaine 45. Pour chaque agent sélectionné, et chaque compétence des agents EF, nous estimons les paramètres des modèles B1, B2, A2 et A3 en nous basant sur toutes les données recueillies jusqu'au jeudi de la semaine 45. (Nous avons omis le modèle A1, car il est dominé par A2 et A3.) En utilisant ces paramètres estimés, nous pouvons générer des temps de service pour chaque agent pour cette journée du vendredi. La sous-section 7.7.1 explique comment les temps de service qui sont de loi log-normale sont générés pour chaque paire (agent i, type d'appel j). Le tableau 7.8 donne les moyennes des temps de service observées pour ce vendredi, les moyennes prédites pour chaque modèle, σ^2 et ${\sigma^2}_\gamma$ pour chaque agent et compétence. Les agents 11 et 12 sont les agents EF. Les tableaux 7.9 et 7.10 donnent le RMSE et le MAPE des erreurs de prédictions pour les types F et E pour chaque modèle.

Notre modèle a été simulé pour $r = 10\,000$ jours indépendants sous chaque modèle de temps de service avec tous les autres paramètres du système inchangés. Nous avons utilisé l'inversion avec des nombres aléatoires communs à travers les quatre modèles pour générer les temps de service de loi log-normale. Ainsi, la seule

Agent	skill	M	$\hat{M}$				σ^2				${\sigma^2}_\gamma$	
			B1	B2	A2	A3	B1	B2	A2	A3	A2	A3
1	F	549	415	563	562	562	1895421	261328	209593	210088	75731	74512
2	F	649	415	500	500	501	1895421	138991	134835	172225	233	6979
3	F	280	415	321	299	284	1895421	189993	107504	113044	2721	1752
4	F	523	415	367	386	395	1895421	145158	128098	115399	776	2885
5	F	538	415	443	442	445	1895421	262364	220645	222747	9040	4842
6	F	427	415	461	422	430	1895421	244324	127759	131373	320	2406
7	F	342	415	369	354	330	1895421	131124	73490	83064	1230	5495
8	F	446	415	480	447	449	1895421	296428	212969	269650	22	5217
9	F	397	415	419	408	414	1895421	134825	105224	90411	4226	2591
10	F	387	415	424	424	424	1895421	200555	202556	186808	4896	5710
11	F	385	415	362	388	401	1895421	149662	126127	129976	4039	41720
12	F	453	415	409	487	499	1895421	105126	127125	130946	4132	42530
11	E	417	412	378	428	415	331230	157662	125483	129976	13921	16347
12	E	444	412	456	456	451	331230	158621	124253	128232	14526	17340

Tableau 7.8 : Moyennes des temps de service observées M et prédites $\hat{M}$, σ^2 et σ^2_γ pour certains agents au vendredi de la semaine 45 pour chaque modèle.

F				
Modeles	B1	B2	A2	A3
RMSE	112.06	78.85	72.38	69.93
MAPE(%)	17.6	11.82	8.46	8.41

Tableau 7.9 : RMSE et MAPE des erreurs de prédictions pour le type d'appel F au vendredi de la semaine 45.

E				
Modeles	B1	B2	A2	A3
RMSE	22.63	28.66	11.49	4.85
MAPE(%)	4	6	2	1

Tableau 7.10 : RMSE et MAPE des erreurs de prédictions pour le type d'appel E au vendredi de la semaine 45.

différence entre les quatre modèles (dans les simulations) sont les moyennes et les variances de la loi log-normale des temps de service. Nous avons calculé le AWT W et le niveau de service $L = \mathrm{SL}(120)$ (la fraction des appels dont l'attente réelle est inférieure ou égale à 120 secondes), pour chaque journée simulée. Le tableau 7.11 rapporte des intervalles de confiance de 95% pour $\mathbb{E}[W]$ et $\mathbb{E}[L]$ sur la base de

ces r simulations, pour chaque modèle et type d'appel. Le tableau 7.11 illustre les différences potentielles dans le SL et le AWT à travers les différents modèles. Par exemple, la différence de la moyenne SL entre les modèles A2 et B2 est proche de 2% et la différence entre le AWT des modèles A2 et B2 est proche de 4%. (comme prévu, les différences de AWT et SL comparées avec le modèle B1 sont encore plus grandes). Bien que de telles différences peuvent apparaître minimes à première vue, elles pourraient conduire à de grandes différences de coûts en pratique. Par exemple, ACS Wireless a trouvé que la diminution du AWT que de 0.6 secondes peut conduire à des économies de 8 million de dollar par an (Hanks, 2014) dans un cas particulier. En outre, de petites différences de pourcentage dans le SL peuvent signifier la différence entre respecter et violer les accords de niveau de service, ce qui peut entraîner des sanctions lourdes pour le centre d'appels. Ainsi, nos résultats numériques montrent que les différents modèles de temps de service peuvent conduire à différentes moyennes de performances du système. Ces différences pourraient potentiellement conduire à des réductions de coûts importantes en pratique. La figure 7.17 affiche les histogrammes des r valeurs de W et L avec chacun des quatre modèles, pour chaque type d'appel. Pour les types types d'appels, nous observons que les distributions de W et de L sont très similaires pour les modèles A2 et A3 (elles ont presque les mêmes fréquences). Leur différence avec B2 est visible, mais elle n'est pas élevée. Par contre avec B1, nous observons une différence beaucoup plus grande.

Nos modèles de temps de service pourraient être utilisés dans la pratique pour permettre une évaluation plus précise de la performance des différents agents. Ceci, à son tour, permet une meilleure classification des agents en groupes qui traitent les différents types d'appels. Dans la sous-section suivante, nous étudions les effets de la sélection des agents.

7.7.3 Impact de la sélection d'agents

Pour illustrer l'impact potentiel de la sélection d'agent pour une journée donnée, nous considérons maintenant des situations supplémentaires pour les modèles B2,

	SL (%)		AWT (s)	
Modèle	F	E	F	E
$B1$	82.68 ± 0.31	56.17 ± 0.41	68.72 ± 1.40	261.80 ± 3.9
$B2$	78.28 ± 0.25	55.31 ± 0.40	73.76 ± 1.30	166.68 ± 2.00
$A2$	79.58 ± 0.23	55.91 ± 0.40	68.20 ± 1.23	160.06 ± 1.98
$A3$	79.26 ± 0.24	54.95 ± 0.40	69.75 ± 1.27	162.42 ± 1.99

Tableau 7.11 : Performances estimées et intervalles de confiance pour notre modèle N.

A2 et A3 (nous omettons le modèle $B1$ car il suppose que tous les agents sont identiques). Dans l'exemple précédent, les agents F nommés 1, 2 et 8 du tableau 7.8 sont lents. Nous remplaçons ces trois agents F lents par trois agents F rapides. Nous supposons que nous avons des doubles des agents rapides 3, 7 et 10, qui remplacent respectivement nos agents lents 1, 2 et 8 dans l'exemple précédent. Comme dans l'exemple précédent, nous avons simulé $r = 10\,000$ jours indépendants et calculé W et L, pour chaque jour. Le tableau 7.12 rapporte des intervalles de confiance à 95% pour $\mathbb{E}[W]$ et $\mathbb{E}[L]$ sur la base de ces r simulations, pour chaque modèle et chaque type d'appel. Nous constatons qu'en comparant avec le tableau 7.11, les performances sont nettement améliorées dans le système, à travers tous les modèles de temps de service. Nous notons également au passage que les différences entre les performances selon des modèles alternatifs sont plus grandes avec ce choix de groupe d'agent. La Figure 7.18 affiche un histogramme des r valeurs de W pour chacun des quatre modèles, pour chaque type d'appel. Contrairement au cas précédent, nous observons ici une différence importante entre les distributions du modèle A2 et A3. Il y a une différence significative entre les modèles alternatifs et B2.

	SL (%)		AWT (s)	
Modèle	F	E	F	E
$B2$	83.25 ± 0.28	59.87 ± 0.37	57.57 ± 1.03	147.17 ± 1.80
$A2$	81.94 ± 0.31	58.38 ± 0.40	60.65 ± 1.11	151.50 ± 1.8
$A3$	85.38 ± 0.26	60.46 ± 0.37	48.89 ± 0.95	143.80 ± 1.69

Tableau 7.12 : Performance estimée et intervalles de confiance pour notre modèle N, avec des agents rapides.

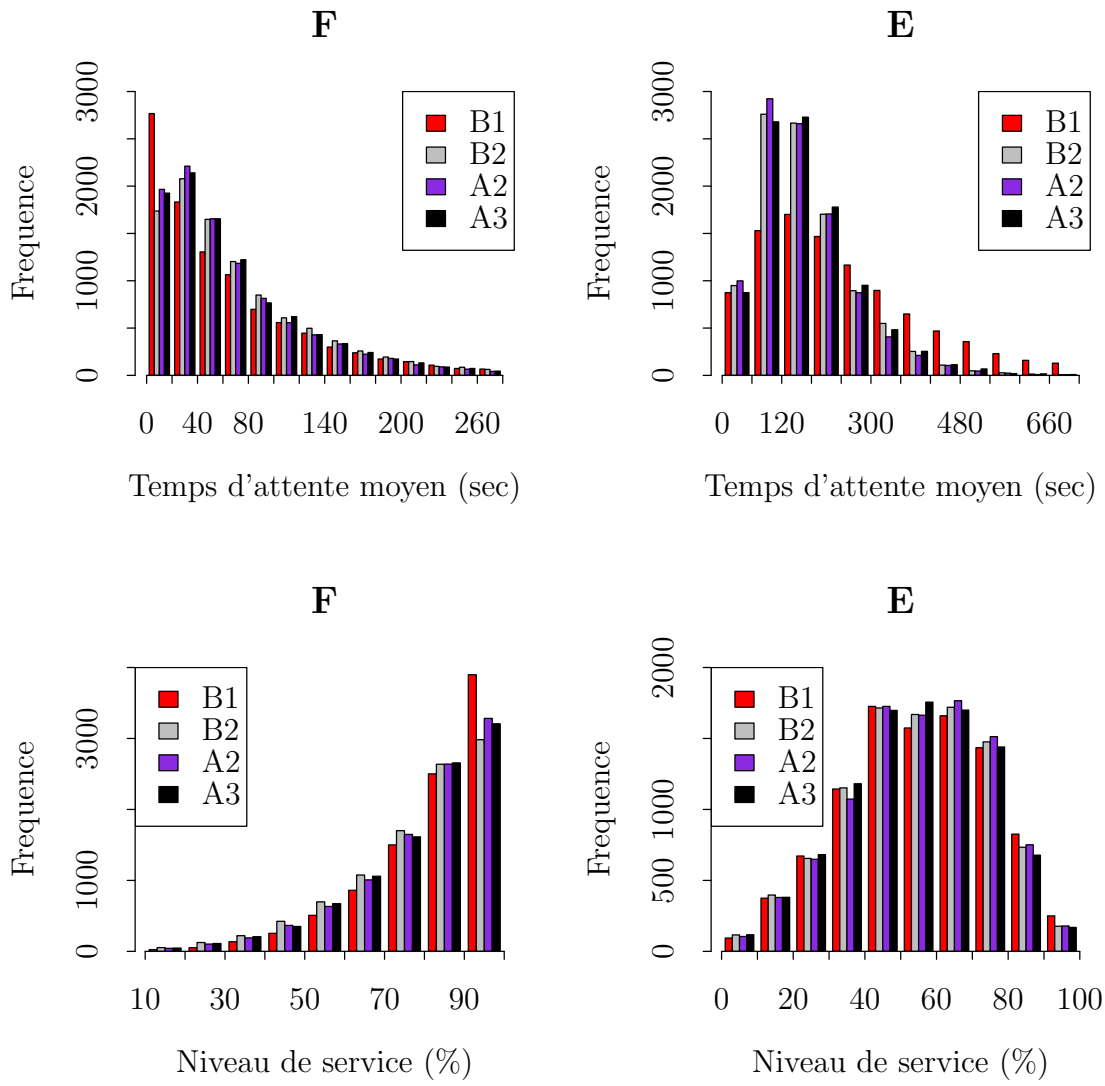

Figure 7.17 : Histogramme du temps service moyen et niveau de service pour les types d'appels 1 et 2 avec tous les modèles.

Supposons maintenant que nous remplaçons l'agent EF numéro 11 par un clone de l'agent 12 (le plus lent) dans notre exemple original, et nous répétons la même expérience. Dans le tableau 7.13, nous présentons les résultats pour ce cas. La comparaison du tableau 7.13 avec le tableau 7.11 montre que les performances du système se sont dégradées de façon significative, en dépit du changement d'un

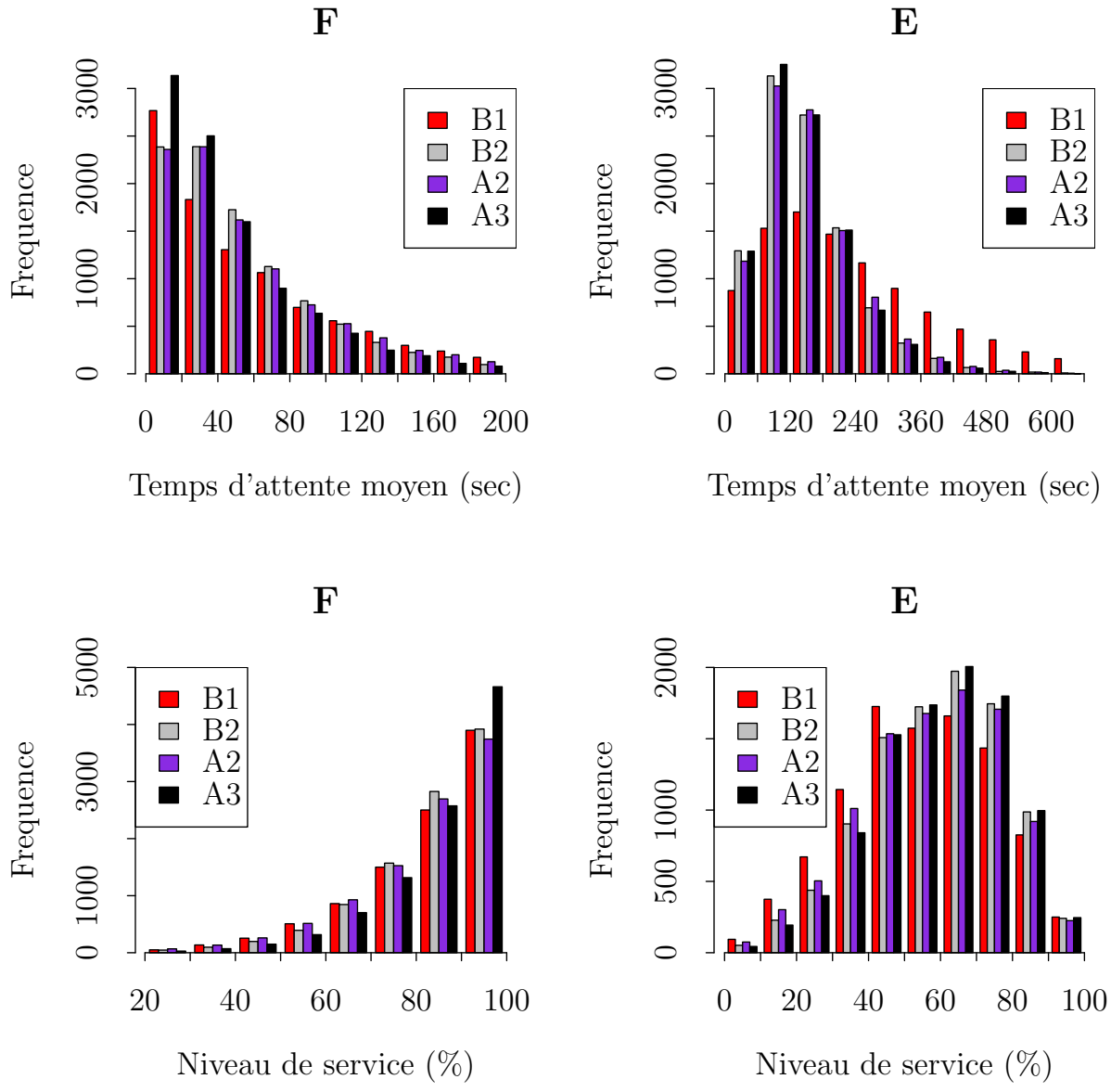

Figure 7.18 : Histogrammes du temps de service moyen et du niveau de service pour les types d'appels 1 et 2 avec tous les modèles.

seul agent. Ainsi, nos résultats numériques montrent que la sélection des groupes d'agents spécifiques avec des vitesses de traitement différentes, basée sur nos modèles de temps de service, peut conduire à des différences significatives dans les performances du système. La figure 7.19 affiche les histogrammes des r valeurs de

W et L avec chacun des quatre modèles, pour chaque type d'appel. Elle montre une différence importante entre les distributions des modèles A2 et A3. Mais cette fois, nous observons que les distributions de B2 et A2 sont très similaires.

	SL (%)		AWT (s)	
Modèle	F	E	F	E
$B2$	76.54 ± 0.36	52.68 ± 0.40	78.89 ± 1.38	175.86 ± 2.02
A2	76.88 ± 0.36	52.68 ± 0.40	78.89 ± 1.38	175.86 ± 2.02
$A3$	76.13 ± 0.26	50.78 ± 0.36	76.48 ± 0.90	182.28 ± 1.65

Tableau 7.13 : Performance estimée et intervalles de confiance pour notre modèle N avec des agents EF lents.

Périodes	Staffing
1	5
2	5
3	5
4	5
5	8
6	8
7	8
8	8
9	10
10	10

Périodes	Staffing
11	10
12	10
13	10
14	10
15	10
16	10
17	9
18	9
19	9
20	9

Périodes	Staffing
21	10
22	10
23	10
24	10
25	10
26	10
27	10
28	10
29	8
20	8

Périodes	Staffing
31	8
32	8
33	8
34	8
35	8
36	8
37	6
38	6
39	6
40	6

Tableau 7.14 : Staffing des périodes les agents F

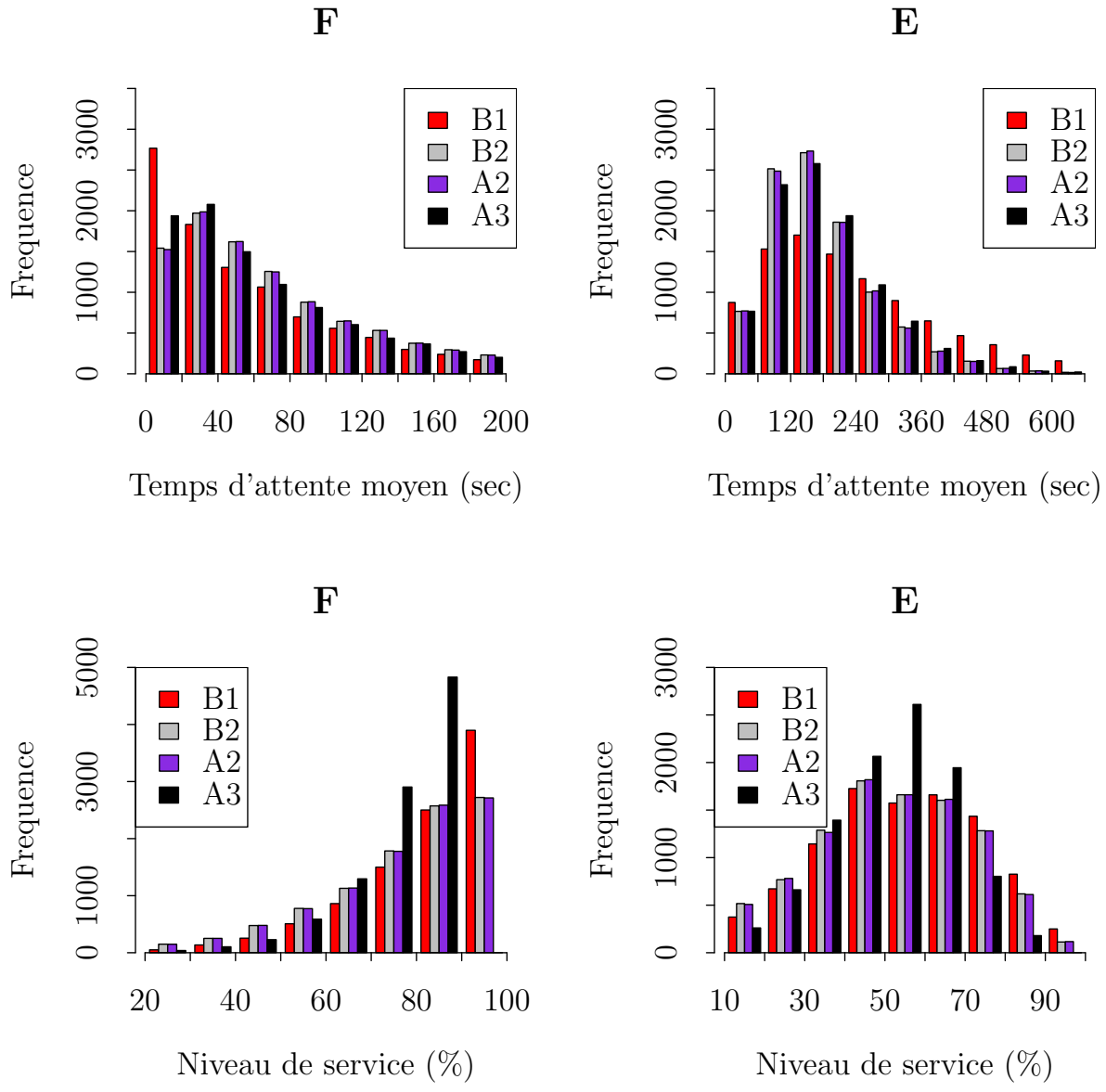

Figure 7.19 : Histogrammes du temps de service moyen et du niveau de service pour les types d'appels 1 et 2 avec tous les modèles.

7.8 Conclusion et remarques

Dans ce chapitre, nous avons adopté une approche basée sur les données pour la modélisation des temps de service dans les centres d'appels. Nous avons évalué la

Périodes	Forme	Échelle
1	17.1	12.5
2	14.1	12.0
3	10.3	14.0
4	11.3	15.5
5	9.4	17.4
6	11.7	18.4
7	28.7	18.8
8	35.8	18.6
9	18.6	20.7
10	35.2	21.3

Periods	Forme	Échelle
11	27.1	22.3
12	37.6	22.9
13	29.9	22.8
14	52.5	21.1
15	42.4	20.2
16	44.8	19.3
17	33.6	17.8
18	15.0	17.1
19	35.1	18.4
20	14.4	18.1

Périodes	Forme	Échelle
21	38.4	20.1
22	18.4	19.5
23	13.9	19.0
24	16.4	19.1
25	16.8	19.5
26	33.9	19.3
27	24.5	19.7
28	36.7	19.2
29	54.3	19.7
30	40.6	19.9

Périodes	Forme	Échelle
31	62.9	18.7
32	36.0	17.4
33	14.6	15.4
34	11.4	14.9
35	8.4	13.7
36	15.1	13.3
37	11.1	11.3
38	15.1	9.8
39	14.1	9.0
40	22.2	7.5

Tableau 7.15 : Paramètres de forme et d'échelle des périodes pour la distribution Gamma dans le processus d'arrivée pour type d'appel F.

qualité de l'ajustement aux données, à la fois dans l'échantillon et hors échantillon, de plusieurs modèles de temps de service. Nos modèles intègrent plusieurs propriétés couramment observées dans la pratique telles que : (1) l'hétérogénéité agent/type d'appel, (2) une performance des agents dépendant du temps, (3) l'existence de corrélations croisées/sérielles dans les données. En général, nous avons constaté que les modèles qui exploitent ces propriétés sont supérieurs aux modèles qui ne le font pas. Pour démontrer l'avantage supplémentaire de cette amélioration de la qualité de l'ajustement, nous avons présenté et discuté les résultats d'expériences de simulation qui ont montré que : (1) la sélection de différents modèles de temps de service, peut avoir un impact significatif sur les performances moyennes du système, pou-

Périodes	Forme	Échelle
1	4.8	0.3
2	5.4	0.3
3	6.2	0.3
4	6.4	0.4
5	8.2	0.5
6	8.8	0.5
7	9.6	0.5
8	10.4	0.5
9	10.6	0.6
10	9.9	0.6

Périodes	Forme	Échelle
11	10.0	0.6
12	14.0	0.6
13	10.0	0.6
14	10.1	0.6
15	10.3	0.5
16	10.4	0.5
17	13.5	0.4
18	11.6	0.4
19	8.9	0.4
20	8.2	0.5

Périodes	Forme	Échelle
21	6.5	0.5
22	10.3	0.6
23	8.7	0.5
24	11.4	0.5
25	10.8	0.5
26	9.7	0.5
27	14.8	0.5
28	11.2	0.5
29	12.1	0.5
30	17.1	0.5

Périodes	Forme	Échelle
31	10.9	0.5
32	10.6	0.5
33	9.2	0.5
34	8.1	0.4
35	8.2	0.4
36	7.2	0.3
37	6.0	0.3
38	6.0	0.2
39	4.2	0.2
40	6.5	0.1

Tableau 7.16 : Paramètres de forme et d'échelle des périodes pour la distribution Gamma dans le processus d'arrivée pour type d'appel E.

vant conduire à d'importante réduction des coûts, et (2) nos modèles de temps de service peuvent être utilisés pour faciliter la classification des agents dans différents groupes, et les performances du système sous les différents groupes peuvent être considérablement différentes. Dans notre étude de simulation, nous avons considéré un centre d'appels relativement petit pour illustrer l'impact opérationnel dû aux différents modèles de temps de service. En réalité, nous prévoyons que cet impact sera plus marqué dans les centres d'appels de taille petite et moyenne (d'une dizaine d'agents), et peut être plus petit, dû à l'effet de moyenne avec une population d'agents très grande. Néanmoins, les propriétés statistiques (des temps de service) que nous avons observées dans notre ensemble de données devraient continuer à

être observées, quelle que soit la taille du centre d'appels.

Periods	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
P1	1.0	0.5	0.2	0.5	0.2	0.2	0.5	0.4	0.3	0.4	0.3	0.3	0.5	0.5	0.4	0.3	0.4	0.2	0.3	0.5
P2	0.5	1.0	0.7	0.8	0.5	0.8	0.6	0.6	0.6	0.7	0.5	0.7	0.7	0.6	0.6	0.6	0.7	0.5	0.7	0.8
P3	0.2	0.7	1.0	0.6	0.6	0.6	0.7	0.6	0.6	0.7	0.5	0.6	0.7	0.6	0.7	0.6	0.7	0.7	0.8	0.7
P4	0.5	0.8	0.6	1.0	0.6	0.6	0.7	0.6	0.7	0.6	0.6	0.7	0.6	0.5	0.6	0.5	0.7	0.6	0.6	0.8
P5	0.2	0.5	0.6	0.6	1.0	0.4	0.7	0.5	0.7	0.6	0.6	0.5	0.5	0.6	0.6	0.7	0.7	0.6	0.7	0.7
P6	0.2	0.8	0.6	0.6	0.4	1.0	0.6	0.6	0.7	0.8	0.4	0.6	0.6	0.5	0.6	0.6	0.5	0.5	0.6	0.7
P7	0.5	0.6	0.7	0.7	0.7	0.6	1.0	0.7	0.7	0.7	0.4	0.6	0.7	0.6	0.8	0.6	0.7	0.6	0.7	0.8
P8	0.4	0.6	0.6	0.6	0.5	0.6	0.7	1.0	0.7	0.5	0.6	0.6	0.6	0.5	0.4	0.5	0.5	0.5	0.5	0.5
P9	0.3	0.6	0.7	0.7	0.7	0.7	0.7	0.7	1.0	0.6	0.5	0.6	0.6	0.5	0.6	0.5	0.6	0.7	0.7	0.7
P10	0.4	0.7	0.7	0.6	0.6	0.8	0.7	0.5	0.6	1.0	0.6	0.8	0.6	0.8	0.7	0.8	0.8	0.5	0.8	0.7
P11	0.3	0.5	0.5	0.6	0.6	0.4	0.4	0.6	0.5	0.6	1.0	0.9	0.7	0.8	0.6	0.9	0.8	0.7	0.8	0.6
P12	0.3	0.7	0.6	0.7	0.6	0.6	0.6	0.6	0.6	0.8	0.9	1.0	0.8	0.9	0.8	0.9	0.9	0.7	0.9	0.8
P13	0.5	0.7	0.7	0.6	0.5	0.6	0.7	0.6	0.6	0.6	0.7	0.8	1.0	0.8	0.8	0.7	0.8	0.7	0.8	0.8
P14	0.5	0.6	0.6	0.5	0.5	0.5	0.6	0.5	0.5	0.8	0.8	0.9	0.8	1.0	0.8	0.9	0.9	0.6	0.9	0.8
P15	0.4	0.6	0.7	0.6	0.6	0.6	0.8	0.4	0.6	0.7	0.6	0.8	0.8	0.8	1.0	0.8	0.9	0.8	0.9	0.8
P16	0.3	0.6	0.6	0.5	0.6	0.6	0.6	0.5	0.5	0.8	0.9	0.9	0.7	0.9	0.8	1.0	0.9	0.7	0.9	0.7
P17	0.4	0.7	0.7	0.7	0.7	0.5	0.7	0.5	0.6	0.8	0.8	0.9	0.8	0.9	0.9	0.9	1.0	0.8	0.9	0.8
P18	0.2	0.5	0.7	0.6	0.7	0.5	0.6	0.5	0.7	0.5	0.7	0.7	0.7	0.6	0.8	0.7	0.8	1.0	0.8	0.8
P19	0.3	0.7	0.8	0.6	0.6	0.6	0.7	0.5	0.7	0.8	0.8	0.9	0.8	0.9	0.9	0.9	0.9	0.8	1.0	0.8
P20	0.5	0.8	0.7	0.8	0.7	0.7	0.8	0.5	0.7	0.7	0.6	0.8	0.8	0.8	0.8	0.8	0.7	0.8	0.8	0.8	1.0
P21	0.4	0.8	0.6	0.5	0.5	0.7	0.6	0.6	0.5	0.7	0.4	0.6	0.6	0.7	0.7	0.6	0.6	0.5	0.7	0.8
P22	0.3	0.6	0.5	0.3	0.4	0.7	0.6	0.4	0.6	0.6	0.4	0.5	0.6	0.5	0.8	0.5	0.6	0.6	0.6	0.6
P23	0.4	0.6	0.5	0.6	0.4	0.5	0.5	0.5	0.6	0.5	0.4	0.5	0.4	0.5	0.6	0.5	0.5	0.5	0.4	0.5
P24	0.0	0.6	0.7	0.6	0.5	0.8	0.7	0.5	0.7	0.7	0.5	0.6	0.6	0.6	0.7	0.6	0.7	0.7	0.7	0.7
P25	0.1	0.6	0.7	0.7	0.7	0.7	0.7	0.5	0.7	0.7	0.6	0.8	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.8
P26	0.1	0.6	0.8	0.5	0.6	0.6	0.6	0.6	0.7	0.7	0.8	0.8	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.6
P27	0.2	0.7	0.7	0.6	0.7	0.7	0.8	0.6	0.6	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.8	0.7	0.8	0.7
P28	0.1	0.5	0.5	0.7	0.6	0.4	0.6	0.4	0.5	0.6	0.8	0.8	0.6	0.7	0.6	0.7	0.9	0.7	0.7	0.6
P29	0.4	0.4	0.3	0.4	0.2	0.3	0.5	0.2	0.0	0.5	0.6	0.6	0.7	0.8	0.6	0.6	0.7	0.4	0.6	0.5
P30	0.5	0.5	0.5	0.5	0.4	0.4	0.5	0.4	0.4	0.7	0.8	0.9	0.8	1.0	0.7	0.8	0.8	0.5	0.8	0.7
P31	0.5	0.4	0.3	0.4	0.4	0.2	0.4	0.3	0.3	0.6	0.8	0.8	0.7	0.9	0.6	0.8	0.8	0.5	0.7	0.5
P32	0.1	0.3	0.3	0.3	0.3	0.2	0.3	0.1	0.2	0.5	0.6	0.5	0.3	0.5	0.4	0.5	0.6	0.4	0.5	0.3
P33	0.2	0.4	0.4	0.6	0.5	0.5	0.7	0.6	0.5	0.5	0.4	0.5	0.4	0.3	0.5	0.4	0.5	0.4	0.4	0.5
P34	0.3	0.6	0.6	0.7	0.6	0.5	0.8	0.6	0.6	0.7	0.5	0.7	0.5	0.6	0.6	0.6	0.8	0.6	0.7	0.7
P35	0.3	0.5	0.6	0.5	0.4	0.5	0.6	0.6	0.6	0.6	0.6	0.7	0.7	0.7	0.6	0.6	0.7	0.6	0.7	0.6
P36	0.5	0.5	0.6	0.5	0.4	0.3	0.6	0.6	0.6	0.6	0.5	0.7	0.8	0.8	0.6	0.6	0.7	0.5	0.7	0.6
P37	0.6	0.7	0.6	0.7	0.5	0.4	0.5	0.6	0.6	0.7	0.6	0.8	0.8	0.8	0.6	0.7	0.8	0.6	0.8	0.7
P38	0.5	0.6	0.4	0.5	0.3	0.2	0.3	0.6	0.4	0.6	0.5	0.6	0.6	0.6	0.3	0.5	0.5	0.4	0.6	0.5
P39	0.5	0.3	0.1	0.2	0.0	-0.1	0.1	0.1	0.1	0.0	0.4	0.4	0.4	0.6	0.3	0.5	0.4	0.2	0.4	0.4
P40	0.7	0.4	0.3	0.4	0.2	0.1	0.3	0.3	0.2	0.6	0.5	0.5	0.5	0.8	0.4	0.6	0.6	0.3	0.6	0.5

Tableau 7.17 : Matrice de corrélation entre les taux d'arrivée pour le type d'appel F partie 1.

Périodes	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40
P1	0.4	0.3	0.4	0.0	0.1	0.1	0.2	0.1	0.4	0.5	0.5	0.1	0.2	0.3	0.3	0.5	0.6	0.5	0.5	0.7
P2	0.8	0.6	0.6	0.6	0.6	0.6	0.7	0.5	0.4	0.5	0.4	0.3	0.4	0.6	0.5	0.5	0.7	0.6	0.3	0.4
P3	0.6	0.5	0.5	0.7	0.7	0.8	0.7	0.5	0.3	0.5	0.3	0.3	0.4	0.6	0.6	0.6	0.6	0.4	0.1	0.3
P4	0.5	0.3	0.6	0.6	0.7	0.5	0.6	0.7	0.4	0.5	0.4	0.3	0.6	0.7	0.5	0.5	0.7	0.5	0.2	0.4
P5	0.5	0.4	0.4	0.5	0.7	0.6	0.7	0.6	0.2	0.4	0.4	0.3	0.5	0.6	0.4	0.4	0.5	0.3	0.0	0.2
P6	0.7	0.7	0.5	0.8	0.7	0.6	0.7	0.4	0.3	0.4	0.2	0.2	0.5	0.5	0.5	0.3	0.4	0.2	-0.1	0.1
P7	0.6	0.6	0.5	0.7	0.7	0.6	0.8	0.6	0.5	0.5	0.4	0.3	0.7	0.8	0.6	0.6	0.5	0.3	0.1	0.3
P8	0.6	0.4	0.5	0.5	0.5	0.6	0.6	0.4	0.2	0.4	0.3	0.1	0.6	0.6	0.6	0.6	0.6	0.6	0.1	0.3
P9	0.5	0.6	0.5	0.7	0.7	0.7	0.6	0.5	0.2	0.4	0.3	0.2	0.5	0.6	0.6	0.6	0.6	0.4	0.1	0.2
P10	0.7	0.6	0.6	0.7	0.7	0.7	0.7	0.6	0.5	0.7	0.6	0.5	0.5	0.7	0.6	0.6	0.7	0.6	0.4	0.6
P11	0.4	0.4	0.4	0.5	0.6	0.8	0.7	0.8	0.6	0.8	0.8	0.6	0.4	0.5	0.6	0.5	0.6	0.5	0.4	0.5
P12	0.6	0.5	0.5	0.6	0.8	0.8	0.8	0.8	0.6	0.9	0.8	0.5	0.5	0.7	0.7	0.7	0.8	0.6	0.4	0.5
P13	0.6	0.6	0.4	0.6	0.7	0.7	0.7	0.6	0.7	0.8	0.7	0.3	0.4	0.5	0.7	0.8	0.8	0.6	0.4	0.5
P14	0.7	0.5	0.5	0.6	0.7	0.7	0.7	0.7	0.8	1.0	0.9	0.5	0.3	0.6	0.7	0.8	0.8	0.6	0.6	0.8
P15	0.7	0.8	0.6	0.7	0.8	0.7	0.7	0.6	0.6	0.7	0.6	0.4	0.5	0.6	0.6	0.6	0.6	0.3	0.3	0.4
P16	0.6	0.5	0.5	0.6	0.7	0.8	0.7	0.7	0.6	0.8	0.8	0.5	0.4	0.6	0.6	0.6	0.7	0.5	0.5	0.6
P17	0.6	0.6	0.5	0.7	0.8	0.8	0.8	0.9	0.7	0.8	0.8	0.6	0.5	0.8	0.7	0.7	0.8	0.5	0.4	0.6
P18	0.5	0.6	0.5	0.7	0.8	0.8	0.7	0.7	0.4	0.5	0.5	0.4	0.4	0.6	0.6	0.5	0.6	0.4	0.2	0.3
P19	0.7	0.6	0.4	0.7	0.8	0.9	0.8	0.7	0.6	0.8	0.7	0.5	0.4	0.7	0.7	0.7	0.8	0.6	0.4	0.6
P20	0.8	0.6	0.5	0.7	0.8	0.6	0.7	0.6	0.5	0.7	0.5	0.3	0.5	0.7	0.6	0.6	0.7	0.5	0.4	0.5
P21	1.0	0.6	0.4	0.7	0.7	0.6	0.5	0.3	0.4	0.5	0.4	0.2	0.5	0.6	0.6	0.5	0.6	0.5	0.3	0.4
P22	0.6	1.0	0.5	0.6	0.7	0.6	0.6	0.4	0.3	0.4	0.3	0.3	0.5	0.4	0.4	0.3	0.4	0.2	0.1	0.1
P23	0.4	0.5	1.0	0.6	0.6	0.4	0.6	0.5	0.1	0.3	0.3	0.4	0.6	0.5	0.5	0.4	0.5	0.4	0.2	0.3
P24	0.7	0.6	0.6	1.0	0.8	0.8	0.8	0.7	0.4	0.5	0.3	0.5	0.7	0.7	0.7	0.5	0.5	0.4	0.1	0.2
P25	0.7	0.7	0.6	0.8	1.0	0.8	0.8	0.8	0.4	0.6	0.4	0.5	0.7	0.7	0.6	0.5	0.6	0.3	0.1	0.2
P26	0.6	0.6	0.4	0.8	0.8	1.0	0.8	0.7	0.5	0.7	0.6	0.6	0.4	0.7	0.6	0.5	0.6	0.4	0.2	0.3
P27	0.5	0.6	0.6	0.8	0.8	0.8	1.0	0.8	0.4	0.6	0.5	0.5	0.6	0.7	0.7	0.5	0.5	0.4	0.1	0.2
P28	0.3	0.4	0.5	0.7	0.8	0.7	0.8	1.0	0.7	0.7	0.7	0.7	0.6	0.7	0.5	0.5	0.6	0.4	0.3	0.4
P29	0.4	0.3	0.1	0.4	0.4	0.5	0.4	0.7	1.0	0.9	0.9	0.6	0.2	0.4	0.4	0.7	0.6	0.4	0.6	0.7
P30	0.5	0.4	0.3	0.5	0.6	0.7	0.6	0.7	0.9	1.0	0.9	0.5	0.3	0.6	0.7	0.9	0.8	0.6	0.7	0.8
P31	0.4	0.3	0.3	0.3	0.4	0.6	0.5	0.7	0.9	0.9	1.0	0.7	0.2	0.5	0.6	0.8	0.7	0.7	0.8	0.9
P32	0.2	0.3	0.4	0.5	0.5	0.6	0.5	0.7	0.6	0.5	0.7	1.0	0.4	0.5	0.3	0.3	0.5	0.5	0.5	0.5
P33	0.5	0.5	0.6	0.7	0.7	0.4	0.6	0.6	0.2	0.3	0.2	0.4	1.0	0.8	0.5	0.3	0.3	0.2	0.0	0.1
P34	0.6	0.4	0.5	0.7	0.7	0.7	0.7	0.7	0.4	0.6	0.5	0.5	0.8	1.0	0.7	0.6	0.6	0.5	0.3	0.5
P35	0.6	0.4	0.5	0.7	0.6	0.6	0.7	0.5	0.4	0.7	0.6	0.3	0.5	0.7	1.0	0.8	0.7	0.5	0.4	0.5
P36	0.5	0.3	0.4	0.5	0.5	0.5	0.5	0.5	0.7	0.9	0.8	0.3	0.3	0.6	0.8	1.0	0.8	0.6	0.7	0.7
P37	0.6	0.4	0.5	0.5	0.6	0.6	0.5	0.6	0.6	0.8	0.7	0.5	0.3	0.6	0.7	0.8	1.0	0.9	0.7	0.8
P38	0.5	0.2	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.6	0.7	0.5	0.2	0.5	0.5	0.6	0.9	1.0	0.8	0.8
P39	0.3	0.1	0.2	0.1	0.1	0.2	0.1	0.3	0.6	0.7	0.8	0.5	0.0	0.3	0.4	0.7	0.7	0.8	1.0	0.9
P40	0.4	0.1	0.3	0.2	0.2	0.3	0.2	0.4	0.7	0.8	0.9	0.5	0.1	0.5	0.5	0.7	0.8	0.8	0.9	1.0

Tableau 7.18 : Matrice de corrélation entre les taux d'arrivée pour le type d'appel F partie 2.

Périodes	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
P1	1.0	0.9	0.9	0.8	0.8	1.0	1.0	1.0	0.9	1.0	0.9	0.9	0.9	1.0	1.0	1.0	0.9	0.9	1.0	1.0
P2	0.9	1.0	0.9	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.9	0.8	0.9	0.9	0.9	0.9
P3	0.9	0.9	1.0	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9
P4	0.8	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.9	0.8	0.9	0.9	0.9	0.9
P5	0.8	0.9	0.9	0.9	1.0	0.9	0.9	0.9	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
P6	1.0	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P7	1.0	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0
P8	1.0	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0
P9	0.9	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P10	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P11	0.9	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	0.9	0.8	0.9	0.9	1.0	0.9	0.9	1.0	1.0	1.0
P12	0.9	0.9	0.9	0.9	0.8	0.9	0.9	1.0	1.0	0.9	1.0	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9
P13	0.9	0.8	0.8	0.8	0.7	0.9	0.9	0.9	0.9	0.8	0.9	1.0	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.9
P14	1.0	0.8	0.8	0.8	0.8	1.0	1.0	1.0	1.0	0.9	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P15	1.0	0.8	0.8	0.8	0.8	1.0	0.9	1.0	1.0	0.9	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P16	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P17	0.9	0.8	0.9	0.8	0.8	1.0	0.9	0.9	1.0	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0
P18	0.9	0.9	0.9	0.9	0.8	0.9	0.9	1.0	1.0	0.9	1.0	0.8	0.9	0.9	0.9	0.9	1.0	1.0	0.9	0.9
P19	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0
P20	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0
P21	1.0	0.8	0.8	0.8	0.8	1.0	0.9	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0
P22	1.0	0.8	0.8	0.8	0.8	0.9	0.9	1.0	1.0	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0
P23	1.0	0.9	0.8	0.8	0.8	1.0	1.0	1.0	1.0	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0
P24	0.9	0.8	0.8	0.8	0.7	0.9	1.0	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.9
P25	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P26	1.0	0.8	0.8	0.8	0.8	1.0	1.0	1.0	1.0	0.9	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P27	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P28	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P29	1.0	0.8	0.8	0.8	0.8	1.0	1.0	1.0	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P30	1.0	0.8	0.8	0.8	0.7	1.0	1.0	1.0	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P31	0.9	0.8	0.8	0.7	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.9
P32	1.0	0.9	0.9	0.8	0.8	1.0	0.9	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P33	1.0	0.8	0.8	0.8	0.8	1.0	0.9	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P34	0.7	0.8	0.8	0.8	0.7	0.8	0.8	0.9	0.9	0.8	0.9	0.8	0.8	0.8	0.7	0.8	0.8	1.0	0.9	0.8
P35	0.9	0.8	0.7	0.7	0.6	0.8	0.9	0.9	0.9	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.9
P36	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.9	1.0	0.9	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
P37	1.0	0.9	0.9	0.9	0.8	1.0	0.9	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0
P38	0.9	0.7	0.8	0.8	0.6	0.9	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
P39	0.9	0.8	0.8	0.8	0.7	0.9	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.9
P40	1.0	0.9	0.9	0.9	0.8	1.0	1.0	1.0	1.0	1.0	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0

Tableau 7.19 : Matrice de corrélation entre les taux d'arrivée pour le type d'appel E partie 1.

Périodes	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40
P1	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.7	0.9	0.9	1.0	0.9	0.9	1.0
P2	0.8	0.8	0.9	0.8	0.9	0.8	0.9	0.9	0.8	0.8	0.8	0.9	0.8	0.8	0.8	0.8	0.9	0.7	0.8	0.9
P3	0.8	0.8	0.8	0.8	0.9	0.8	0.9	0.9	0.8	0.8	0.8	0.9	0.8	0.8	0.7	0.9	0.9	0.8	0.8	0.9
P4	0.8	0.8	0.8	0.8	0.9	0.8	0.9	0.9	0.8	0.8	0.7	0.8	0.8	0.8	0.7	0.9	0.9	0.8	0.8	0.9
P5	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.7	0.6	0.8	0.8	0.6	0.7	0.8
P6	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.8	0.9	1.0	0.9	0.9	1.0
P7	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	0.9	1.0	0.8	0.8	1.0
P8	0.9	0.9	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	1.0
P9	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.9	1.0	1.0	0.9	0.9	1.0	1.0	0.9	0.9	1.0
P10	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.9	0.9	1.0	1.0	0.9	0.9	1.0
P11	0.9	0.9	0.9	0.9	1.0	0.9	1.0	1.0	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.9	1.0	0.9	0.9	1.0
P12	0.9	0.9	0.9	0.8	1.0	0.9	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.9
P13	1.0	0.9	0.9	0.9	1.0	0.9	0.9	0.9	0.9	0.9	0.9	1.0	1.0	0.8	0.8	0.9	0.9	0.9	0.9	0.9
P14	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	1.0	0.9	0.9	1.0
P15	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.7	0.9	1.0	1.0	0.9	0.9	1.0
P16	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	1.0	0.9	0.9	1.0
P17	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	0.9	0.9	0.9	1.0
P18	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.9	1.0	0.8	0.9	0.9	0.9	0.8	0.9
P19	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.9	0.9	1.0	1.0	0.9	0.9	1.0
P20	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	1.0	0.9	0.9	1.0
P21	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.9	0.9	1.0	0.9	0.9	1.0
P22	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	0.9	0.9	0.9	1.0
P23	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	0.9	1.0	0.9	0.9	1.0
P24	0.9	0.9	1.0	1.0	0.9	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.7	0.8	0.8	0.9	0.8	0.7	0.9
P25	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	1.0	0.9	0.9	1.0
P26	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.9	0.9	1.0	0.9	0.9	1.0
P27	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.9	1.0	1.0	0.9	0.9	1.0
P28	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.9	0.9	1.0	1.0	0.9	0.9	1.0
P29	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.9	0.9	0.9	0.9	0.9	1.0
P30	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	0.9	0.9	0.9	0.9	1.0
P31	1.0	0.9	0.9	0.9	0.9	1.0	1.0	0.9	1.0	0.9	1.0	0.9	1.0	0.8	0.9	0.9	0.9	0.9	0.8	0.9
P32	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	1.0	0.9	0.9	1.0
P33	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.9	1.0	0.9	0.9	0.9	1.0
P34	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.9	0.9	0.8	0.8	0.8	0.8	1.0	0.8	0.8	0.7	0.7	0.7	0.8
P35	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	1.0	0.8	0.8	0.9	0.8	0.9
P36	0.9	1.0	0.9	0.8	1.0	0.9	1.0	1.0	0.9	0.9	0.9	1.0	1.0	0.8	0.8	1.0	1.0	1.0	1.0	1.0
P37	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0	0.9	0.9	0.9	1.0	0.9	0.7	0.8	1.0	1.0	0.9	0.9	1.0
P38	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.7	0.9	1.0	0.9	1.0	1.0	0.9
P39	0.9	0.9	0.9	0.7	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.7	0.8	1.0	0.9	1.0	1.0	0.9
P40	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.8	0.9	1.0	1.0	0.9	0.9	1.0

Tableau 7.20 : Matrice de corrélation entre les taux d'arrivée pour le type d'appel E partie 2.

CHAPITRE 8

CONCLUSION

Les systèmes de service de la vie réelle sont souvent très complexes. Il y a beaucoup de sources d'incertitudes qui rendent leur analyse complexe et difficile. Dans cette thèse, nous nous sommes intéressés à deux problèmes dans les centres d'appels multi-compétences. Le premier est la prédiction du temps d'attente des clients dès leur arrivée. Le second est la modélisation des durées de service. Pour chacun des problèmes, nous avons proposé des solutions dont les tests ont donné des résultats très satisfaisants (car les performances obtenues avec nos nouvelles solutions sont largement meilleures que celles des solutions déjà existantes). Dans cette conclusion, nous résumons les contributions principales de cette thèse et proposons des pistes de recherche.

Nous avons proposé des prédicteurs qui utilisent les algorithmes d'apprentissage machine (RS, SK, ANN) pour les systèmes multi-compétences au chapitre 4. Nos prédicteurs retournent une estimation ponctuelle du temps d'attente basée sur une approximation de l'espérance conditionnelle du temps d'attente conditionnelle à l'état actuel du système quand le client entre dans la file. Cet état actuel est représenté par un vecteur d'information (entrée) $\boldsymbol{x}$ qui est défini pour chaque type d'appel k par la longueur de la file d'attente, les longueurs des files des types servis par les mêmes agents, la période d'arrivée de l'appel, et du vecteur de staffing des groupes. Dans nos expériences numériques, pour les systèmes multi-compétences, les trois prédicteurs RS, SK et ANN sont beaucoup plus précis que la meilleure stratégie existante que nous connaissons dans la littérature, LES, qui retourne le temps d'attente du dernier client du même type qui a commencé son service. Pour les systèmes à file d'attente unique avec des agents homogènes et temps de service exponentiels (FILE 1), les prédicteurs RS, SK et ANN ont une précision très similaire à celle du prédicteur optimal QL. Dans les systèmes à file uniques (cas des centres d'appels) les plus réalistes avec des agents hétérogènes et des durées de

service de loi log-normale (FILE 2), les nouveaux prédicteurs sont largement plus précis que QL.

Au chapitre 5, nous étendons la famille de prédicteurs de délais DH en introduisant deux nouveaux prédicteurs de délai, basés sur des heuristiques simples. Le premier prédicteur (E-LES) exploite l'information de délais plus récente, mais incomplète des clients toujours en attente dans la file d'attente . Leurs temps d'attente finaux sont estimés à l'aide d'une simple extrapolation de leur progression dans la file d'attente. L'autre prédicteur (AvgC-LES) une version empirique de la formule QL dans le contexte des systèmes multi-compétences, en utilisant des données historiques. Pour chaque taille de file d'attente, une espérance conditionnelle des temps d'attente est estimée à partir des délais passés de clients qui ont trouvé la même longueur de file d'attente devant eux quand ils sont arrivés. Dans les systèmes FILE 1, nos nouveaux prédicteurs sont meilleurs que les autres prédicteurs DH que nous connaissons, et en plus nous observons que AvgC-LES est très proche du prédicteur optimal QL. Dans les systèmes FILE 2, E-LES et AvgC-LES en plus d'être meilleurs que les autres prédicteurs DH, sont largement plus précis que QL. Pour les systèmes multi-compétence plus réalistes, qui ont généralement des taux d'arrivée et des staffing variables dans le temps, E-LES et AvgC-LES performent également mieux que les autres prédicteurs DH. Bien qu'ils ne battent pas les méthodes de l'apprentissage machine, leur avantage est qu'ils sont plus simples à mettre en œuvre, ont peu de paramètres, et ne nécessitent aucune phase d'entraînement. Ils représentent des alternatives simples intéressantes aux prédicteurs plus complexes.

Enfin, au chapitre 7, nous avons proposé des modèles pour les temps de service qui intègrent plusieurs propriétés observées en pratique dans les données. Les propriétés prises en compte dans les modèles sont : l'hétérogénéité des agents, la dépendance du temps des durées de service, et les corrélations sérielles et croisées entre les temps de service d'un même agent. Les nouveaux modèles s'ajustent mieux aux données que les autres modèles qui n'intègrent pas ces propriétés. Pour démontrer l'avantage supplémentaire de cette amélioration de la qualité de l'ajustement,

nous avons présenté et discuté les résultats d'expériences de simulation qui ont montré que : (i) la sélection de différents modèles de temps de service peut avoir un impact significatif sur les performances moyennes du système, pouvant conduire à d'importantes réductions des coûts, (ii) nos modèles de temps de service peuvent être utilisés pour faciliter la classification des agents dans différents groupes, et les performances du système sous les différents groupes peuvent être considérablement différentes. Ces travaux sur les durées de service ont été faits en collaboration et publiés dans Ibrahim et al. (2016b).

Plusieurs travaux futurs méritent d'être considérés. Pour la prédiction de délai, nous pensons qu'il serait intéressant d'étudier les effets des annonces de délai (estimés par les prédicteurs proposés dans cette thèse) sur le comportement des clients et les performances du système.

Il serait aussi intéressant d'étudier les performances des nouveaux prédicteurs proposés dans un centre d'appels réel. Pour les prédicteurs qui utilisent l'apprentissage machine, cela pourrait se faire à travers des expériences avec les données du système de la vie réelle pour lesquelles toutes les informations nécessaires à la mise en œuvre des prédicteurs sont disponibles (entrée $\boldsymbol{x}$ pour chaque client et sortie y qui est le temps d'attente réellement observé). Pour les prédicteurs DH simples à mettre en oeuvre, nous pouvons les implémenter dans le système et observer leurs performances à la fin de la journée.

Au chapitre 6, nous avons présenté des idées pour adapter les prédicteurs QL dans les centres d'appels multi-compétences. Nous proposons une représentation alternative du centre d'appels multi-compétences en K modèles de files d'attente indépendantes où K est le nombre de types d'appels du centre d'appels. Ainsi pour chaque type d'appel, nous aurons un centre d'appels indépendant avec un seul type d'appel et un groupe d'agents qui traite les appels. La principale difficulté est de trouver des méthodes efficaces qui permettent d'estimer le nombre d'agents pour chacun des modèles afin d'avoir l'équivalence des deux systèmes. Nous avons déjà donné quelques idées, mais il est nécessaire d'étudier leurs efficacités et aussi éventuellement chercher d'autres méthodes.

Nous croyons également qu'il serait intéressant de développer des méthodes qui prédisent et annoncent non pas une estimation ponctuelle du temps d'attente (une estimation de l'espérance), mais plutôt une estimation de la distribution conditionnelle du délai ou au moins certains de ses quantiles. L'annonce de cette distribution va permettre au client de prendre une décision plus éclairée, car les informations qui sont à sa disposition vont être beaucoup plus complètes.

Pour la modélisation des durées de service, une direction possible de recherche future est d'envisager des modèles alternatifs similaires qui intègrent les effets aléatoires, journaliers, ou intra-journaliers lors de la modélisation des temps de service individuels. Pour ce faire, il faut avoir accès à un ensemble de données appel par appel détaillé. Compte tenu des résultats obtenus à la section 7.6, nous prévoyons que ces modèles conduisent à des prévisions plus précises de la future moyenne des temps de service dans le système. Avec un ensemble de données appel par appel détaillé, il serait également possible de tester la précision de la qualité de l'ajustement et prédictive de ces modèles au-delà du temps de service moyen. Autrement dit, nous pourrons tester dans quelle mesure ces modèles s'ajustent à la totalité des distributions des temps de services individuels dans le système. En effet, les décisions opérationnelles complexes dans les centres d'appels vont compter sur des modèles qui s'ajustent et prédisent avec précision ces distributions.

En fait, comme on avait déjà discuté, il faudrait modéliser directement les paramètres de la loi de probabilité de la durée de service de chaque agent, qui peuvent être aléatoires et évoluer dans le temps. Dans nos données, nous observons les moyennes des durées de service de chaque agent par journée. Sachant que la distribution log-normale s'ajuste mieux aux données des durées du service dans les centres d'appels alors il est préférable de modéliser les paramètres d'échelle κ et de forme σ de cette distribution à partir des moyennes observées. Ce serait beaucoup plus naturel et pratique.

BIBLIOGRAPHIE

Software Advice. You need to offer callback : Here are 3 ways to get it, 2014. `http://hello-operator.softwareadvice.com/3-ways-to-offer-callback-0514/`, Accessed June 2016.

O. Z. Akşin, M. Armony et V. Mehrotra. The modern call center : A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007a.

O. Z. Akşin, F. Karaesmen et E. L. Ormeci. A review of workforce cross-training in call centers from an operations management perspective. Dans *In Workforce Cross Training Handbook, D. Nembhard (ed.)*. CRC Press, 2007b.

S. Aldor-Noiman, P. Feigin et A. Mandelbaum. Workload forecasting for a call center : Methodology and a case study. *The Annals of Applied Statistics*, 3(4): 1403–1447, 2009.

E. Ang, S. Kwasnick, M. Bayati, E. Plambeck et M. Aratow. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management*, 18(1):141–156, 2016.

B. Ankenman, B. L. Nelson et J. Staum. Stochastic kriging for simulation metamodeling. *Operations Research*, 58(2):371–382, 2010.

M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51:287–329, 2005.

M. Armony et C. Maglaras. Contact centers with a call-back option and real-time delay information. *Operations Research*, 52:527–545, 2004a.

M. Armony et C. Maglaras. On customer contact centers with a call-back option : Customer decisions, routing rules, and system design. *Operations Research*, 52 (2), 2004b.

M. Armony et A. Mandelbaum. Routing and staffing in large-scale service systems : The case of homogeneous impatient customers and heterogeneous servers. *Operations Research*, 59(1):50–65, 2011.

M. Armony, N. Shimkin et W. Whitt. The impact of delay announcements in many-server queues with abandonments. *Operations Research*, 57:66–81, 2009.

M. Armony et A. Ward. Fair dynamic routing in large-scale heterogeneous server systems. *Operations Research*, 58(3):624–637, 2010.

A. N. Avramidis, W. Chan, M. Gendreau, P. L'Ecuyer et O. Pisacane. Optimizing daily agent scheduling in a multiskill call centers. *European Journal of Operational Research*, 200(3):822–832, 2010.

A. N. Avramidis et P. L'Ecuyer. Modeling and simulation of call centers. Dans M. E. Kuhl, N. M. Steiger, F. B. Armstrong et J. A. Joines, éditeurs, *Proceedings of the 2005 Winter Simulation Conference*, pages 144–152. IEEE Press, 2005.

Y. Bengio, A. C. Courville et P. Vincent. Unsupervised feature learning and deep learning : A review and new perspectives. *CoRR*, abs/1206.5538, 2012. URL `http://arxiv.org/abs/1206.5538`.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn et L. Zhao. Statistical analysis of a telephone call center : A queueing-science perspective. *Journal of American Statistical Association*, 469(100):36–50, 2005.

E. Buist. *Simulation des centres de contacts*. Thèse de doctorat, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada, 2009.

E. Buist et P. L'Ecuyer. A Java library for simulating contact centers. Dans M. E. Kuhl, N. M. Steiger, F. B. Armstrong et J. A. Joines, éditeurs, *Proceedings of the 2005 Winter Simulation Conference*, pages 556–565. IEEE Press, 2005.

W. Chan. Optimisation stochastique pour l'affectation du personnel polyvalent dans un centre d'appels téléphoniques. Mémoire de maîtrise, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada, 2006.

W. Chan. *Optimisation des horaires des agents et du routage des appels dans les centres d'appels*. Thèse de doctorat, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada, 2013.

W. Chan, G. Koole et P. L'Ecuyer. Dynamic call center routing policies using call waiting and agent idle times. *Manufacturing & Service Operations Management*, 16(4):544–560, 2014.

B. Cleveland et J. Mayben. Call center management on fast forward : Succeeding in today's dynamic inbound environment. *Call Center Press*, 1999.

A. F. Colladon, M. Naldi et M. M. Schiraldi. Quality management in the design of TLC call centres. *International Journal of Engineering and Business Management*, 48(5):1–9, 2013.

CRTC. Final standards for quality of service indicators for use in telephone company regulation and other related matters, 2000. Canadian Radio-Television and Telecommunications Commission, Decision CRTC 2000-24. See `http://www.crtc.gc.ca/archive/ENG/Decisions/2000/DT2000-24.htm`.

C. de Boor. *A Practical Guide to Splines*. Numbéro 27 dans Applied Mathematical Sciences Series. Springer-Verlag, New York, 1978.

M. Delasay, A. Ingolfsson et B. Kolfal. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research*, pages 1–19, 2016.

A. Deslauriers. Modélisation et simulation d'un centre d'appels téléphoniques dans un environement mixte. *Master's thesis, Dept. Computer Science and Operations Research, University of Montreal, Montreal, Canada.*, 2003.

S. Ding. *Workforce Management in Call Centers : Forecasting, Staffing and Empirical studies*. Thèse de doctorat, Vrije Universiteit Amsterdam, 2016.

G. Dobson et J. Pinker. The value of sharing lead time information. *IIE Transactions*, 38:171–183, 2006.

J. Dong, P. Feldman et G. Yom-Tov. The impact of dependent service times on large-scale service systems. *Manufacturing and Service Operations Management*, 14(2):262–278, 2012.

J. Dong, P. Feldman et G. Yom-Tov. Service systems with slowdowns : Potential failures and proposed solutions. *Operations Research*, 63(2):305–324, 2015.

J. Dong, E. Yom Tov et G. Yom Tov. The impact of delay announcements on hospital network coordination and waiting times. 2016. Working paper.

L. Dubé-Rioux, B. Schmitt et F. Leclerc. Consumers reactions to waiting : When delays affect the perception of service quality. *Advances in Consumer Research*, 16:59–63, 1989.

S. Dudin, C. Kim, O. Dudina et J. Baek. Queueing system with heterogeous customers as a model of call center with a call-back for lost customers. *Mathematical Problems in Engineering*, 2013:1–13, 2013.

P. Feldman, J. Li, G. Yom-Tov et E. Yom-Tov. Service systems with slowdowns : Potential failures and proposed solutions. *Operations Research*, 63(2):305–324, 2015.

G. M. Dancik. *Maximum Likelihood Estimates of Gaussian Processes*, 2015. URL `https://cran.r-project.org/web/packages/mlegp/index.html`.

N. Gans, G. Koole et A. Mandelbaum. Telephone call centers : Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5: 79–141, 2003.

N. Gans, N. Liu, A. Mandelbaum, H. Shen et H. Ye. Service times in call centers : Agent heterogeneity and learning with some operational consequences. 6:99–123, 2010.

N. Gans et Y. Zhou. A call-routing problem with service-level constraints. *Operations Research*, 51(2):255–271, 2003.

O. Garnett et A. Mandelbaum. An introduction to skill-based routing and its operational complexities. Manuscript, Technion, Israel, 2000.

X. Glorot, A. Bordes et Y. Bengio. Deep sparse rectifier neural networks. Dans J. G. Geoffrey et D. B. Dunson, éditeurs, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, volume 15, pages 315–323. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011. URL `http://www.jmlr.org/proceedings/papers/v15/glorot11a/glorot11a.pdf`.

I. Goodfellow, D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, J. Bergstra, F. Bastien et Y. Bengio. Pylearn2 : A machine learning research library. *arXiv preprint arXiv :1308.4214*, 2013. URL `http://arxiv.org/abs/1308.4214`.

P. Guo et P. Zipkin. Analysis and comparison of queues with different levels of delay information. *Management Sci*, 53:962–970, 2007.

I. Gurvich et W. Whitt. Queue-and-idlness-ratio controls in many-server service systems. *Mathematics of Operations Research*, 34(2):363–396, 2009.

S. Halfin et W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 1981.

J. Hanks. How the right headset affects call center productivity and the bottom line. 2014. URL `http://telecom.hellodirect.com/docs/Tutorials/Productivity.1.080701.asp`.

R. Hassin. Consumer information in markets with random product quality : The case of queues and balking. *Econometrica*, 54:1185–1195, 1986.

M. Hui et D. Tse. What to tell customers in waits of different lengths : an integrative model of service evaluation. *Journal of Marketing*, 60:81–90, 1996a.

M. Hui et D. Tse. What to tell customers in waits of different lengths : An integrative model of service evaluation. *Journal of Marketing*, 60:81–90, April 1996b.

R. Ibrahim, M. Armony et A. Bassamboo. Does the past predict the future ? The case of delay announcements in service systems. *Management Science*, 2016a. Forthcoming.

R. Ibrahim, P. L'Ecuyer, H. Shen et M. Thiongane. Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *European Journal of Operational Research*, 250:480–492, 2016b.

R. Ibrahim et W. Whitt. Real-time delay estimation based on delay history. *MSOM*, Article in Advance:1–19, 2008.

R. Ibrahim et W. Whitt. Real-time delay estimation based on delay history. *Manufacturing and Services Operations Management*, 11:397–415, 2009a.

R. Ibrahim et W. Whitt. Real-time delay estimation based on delay history. *Manufacturing and Services Operations Management*, 11:397–415, 2009b.

R. Ibrahim et W. Whitt. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science*, 55:1729–1742, 2009c.

R. Ibrahim et W. Whitt. Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Production and operation Management*, pages 1–14, 2010.

R. Ibrahim et W. Whitt. Waiting-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, 59:1106–1118, 2011.

O. Jouini, Z. Akşin et Y. Dallery. The impact of delaying the delay announcements. *Operations Research*, 59:1198–1210, 2011a.

O. Jouini, Y. Dallery et Z. Akşin. Modeling call centers with delay information. *Manufacturing and Service Operations Management*, 13:534–548, 2011b.

K. Katz, B. Larson et R. Larson. Prescription for the waiting in line blues : Entertain, enlighten and engage. *Sloan Management Review*, pages 44–53, Winter 1991.

K. L. Katz, B. M. Larson et R. C. Larson. Prescription for the waiting-in-line blues : Entertain, enlighten, and engage. *Sloan Management Review*, pages 44–54, 1999.

T. Kim, P. Kenkel et B. W. Brorsen. Forecasting hourly peak call volume for a rural electric cooperative call center. *Journal of Forecasting*, 31:314–329, 2012.

G. Koole. *Call Center Optimization*. MG books, Amsterdam, 2013.

G. Koole et A. Mandelbaum. Queueing models of call centers : An introduction. *Annals of Operations Research*, pages 41–59, 2002.

Y. LeCun, Y. Bengio et G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. URL `http://dx.doi.org/10.1038/nature14539`.

Y. Liu et W. Whitt. A fluid approximation for the $g(t)/gi/s(t) + gi$ queue, 2010. `http://columbia.edu/www2040`.

Y. Liu et W. Whitt. Large-time asymptotics for the $g_t/m_t/s_t + gi_t$ many-server fluid queue with abandonment. *Queueing Systems*, 67:145–182, 2011.

C. Maglaras et J. Mieghem. Queueing systems with leadtime constraints : A fluid-model approach for admission and sequencing control. *European Journal of Operational Research*, 167:179–207, 2004.

D. Maister. Psychology of waiting lines. harvard business school cases. pages 71–78, 1984.

A. Mandelbaum, W. Massey, M. Reiman et A. Stolyar. Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. *Proceedings of the 37th Allerton Conference*, pages 1095–1104, 1999.

A. Mandelbaum et S. Zeltyn. Service engineering : Data-based course development and teaching. *INFORMS Transactions on Education*, 11(1):3–19, 2011.

V. Mehrotra, K. Ross, G. Ryder et Y. P. Zhou. Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing and Service Operations Management*, 14(1):66–81, 2012.

A. M. Mood, F. A. Graybill et D. C. Boes. *Introduction to the Theory of Statistics*. McGraw Hill, New York, 3th édition, 1974.

T. Morton et A. Vepsalainen. Priority rules and leadtime estimation for job shop scheduling with weighted tardiness costs. *Management Sci*, 33:1036–1047, 1987.

J. Mowen, J. Licata et J. McPhail. Waiting in the emergency room : How to improve patient satisfaction. *Journal of Health Care Marketing*, 16(2):26–33, 1993.

N. Munichor et A. Rafaeli. Numbers or apologies? customer reactions to telephone waiting time fillers. *J. Applied Psychology*, 92(2):511–511–8, 2007.

E. Nakibly. Predicting waiting times in telephone service systems. Mémoire de maîtrise, Technion, Haifa, Israel, 2002.

B. Oreshkin, N. Régnard et P. L'Ecuyer. Rate-based daily arrival process models with application to call centers. *Operations Research*, 64(2):510–527, 2016.

M. Ornek et P. Collier. The determination of in-process inventory and manufacturing lead time in multi-stage production systems. *International J. Oper. and Production Management*, 8:74–80, 1988.

J. Pichitlamken, A. Deslauriers, P. L'Ecuyer et A. N. Avramidis. Modeling and simulation of a telephone call center. Dans *Proceedings of the 2003 Winter Simulation Conference*, pages 1805–1812. IEEE Press, 2003.

M. Pinedo, S. Seshadri et J. G. Shanthikumar. Call centers in financial services : Strategies, technologies, and operations. *Operations and Technologies*, 1999.

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

S. Salcedo-Sanz, M. Naldi, A. M. Pérez-Bellido, J. A. Portilla-Figueras et E. G. Ortíz-García. Evolutionary optimization of service times in interactive voice response systems. *IEEE Transactions on Evolutionary Computation*, 14(4):602–617, 2010.

A. Senderovich, M. Weidlich, A. Gal et A. Mandelbaum. Queue mining for delay prediction in multi-class service processes. *Information Systems*, 53:278–295, 2015.

J. Shanthikumar et U. Sumita. Approximations for the time spent in a dynamic job shop with applications to due date assignment. *International J. Production Research*, 26:1329–1352, 1988.

H. Shen et L. Brown. Nonparametric modelling of time-varying customer service times at a bank call center. *Applied Stochastic Models in Business and Industry*, 22(3):297–311, 2006.

J. Staum. Better simulation metamodeling : the why, what, and how of stochastic kriging. *Winter Simulation Conference*, pages 119–133, 2009.

S. Taylor. Waiting for service : the relationship between delays and evaluations of service. *Journal of Marketing*, 58:56–69, 1994a.

S. Taylor. Waiting for service : The relationship between delays and evaluations of service. *Journal of Marketing*, 58(2):56–69, 1994b.

M. Thiongane, W. Chan et P. L'Ecuyer. Waiting time predictors for multiskill call centers. Dans *Proceedings of the 2015 Winter Simulation Conference*, pages 3073–3084. IEEE Press, 2015.

M. Thiongane, W. Chan et P. L'Ecuyer. New history-based delay predictors for service systems. Dans *Proceedings of the 2016 Winter Simulation Conference*. IEEE Press, 2016.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 7(0):267–288, 1999.

W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45(2):192–207, 1999a.

W. Whitt. Predicting queueing delays. *Management Science*, 45:870–888, 1999b.

W. Whitt. *Stochastic-Process Limits : An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York, 2002.

W. Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10):1449–1461, 2004.

W. Whitt. Fluid models for many-server queues with abandonments. *Operations Research*, 54:37–54, 2006.

S. N. Wood. *Generalized Additive Models : An Introduction with R*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 978-1-58488-474-3 ; 1-58488-474-6.