4e Conférence Francophone de MOdélisation et SIMulation
"Organisation et Conduite d'Activités dans l'Industrie et les Services"
MOSIM'03 – du 23 au 25 avril 2003 - Toulouse (France)

ANALYSE DU PROBLÈME DES RAPPELS ET DIMENSIONNEMENT DANS UN CENTRE D'APPELS

Mohamed Salah AGUIR

Laboratoire Génie Industriel
École Centrale Paris
Grande Voie des Vignes,
92295 Châtenay-Malabry Cedex, France
Mél : salah@lgi.ecp.fr

Fikri KARAESMEN

Department of Industrial Engineering
Koç University
Rumeli Feneri Yolu
80910 Sariyer, Istanbul, Turkey
Mél : fkaraesmen@ku.edu.tr

Zeynep AKSIN

Graduate School of Business
Koç University
80910 Sariyer, Istanbul, Turkey
Mél : zaksin@ku.edu.tr

Fabrice CHAUVET

Bouygues Telecom R&D
10 rue Paul Dautier,
F-78944 VELIZY Cedex, France
Mél : fchauvet@bouyguestelecom.fr

RÉSUMÉ : *Cet article porte sur l'analyse de l'effet des rappels sur la planification dans un centre d'appels téléphoniques. Le problème du renouvellement d'appels est modélisé par une chaîne de Markov à deux dimensions dont l'étude conduit à la distinction du taux de rappel enregistré à partir du taux d'appel observé par le système. Cette étude mène à une nouvelle approche lors du dimensionnement du centre d'appels et elle sera comparée à ce que donne un système M/M/C/K + M qui ne considère pas le phénomène de rappel.*

MOTS-CLÉS : *Rappels, abandons, File M/M/C/K + M, chaîne de Markov, théorie des files d'attente, dimensionnement.*

1. INTRODUCTION

Les centres d'appels téléphoniques représentent une industrie en nette progression. En 2002, d'après le cabinet Cesmo (cartographie des centres d'appels 2002), elle enregistre une croissance annuelle en France de 8,4 % pour passer à 183 000 salariés pour un total de 2900 centres d'appels de plus de 10 positions. Beaucoup de gros centres d'appels ont vu le jour ces dernières années après que les entreprises se soient rendue compte de l'importance des services et du profit que peut engendrer un centre d'appels lorsqu'il est performant. En particulier dans le domaine des télécommunications où les clients peuvent s'inscrire à de nouveaux services directement par l'intermédiaire du centre d'appels. Ceci implique un enjeu important lors de la planification du centre d'appels puisqu'un client qui ne réussit pas à joindre un conseiller de clientèle détériore la qualité de service que doit offrir le centre. Pour remédier à ce problème, il faut savoir bien optimiser le système puisqu'un surdimensionnement est également synonyme de surcoût. C'est d'autant plus délicat avec la taille grandissante des centres d'appels constitués dans de nombreux cas de plusieurs centaines de conseillers de clientèle répondants à des dizaines de milliers d'appels par jour. Ce problème de planification ressemble, donc, de plus en plus à un problème de planification industriel avec la taille des gros centres d'appels.

Le travail décrit dans cet article fait suite à une problématique énoncée par l'opérateur de téléphonie mobile *Bouygues Telecom*. Nous y avons abordé le problème de rappels téléphoniques en nous basant sur un modèle stochastique de chaîne de Markov à deux dimensions. Ce modèle est semblable à un autre étudié par (Tran-Gia and Mandjes, 1997) dans les réseaux de téléphonie mobile mais, contrairement à eux, le nombre des clients dans notre modèle n'est pas limité et ils peuvent patienter dans une file d'attente de capacité non nulle. Cet article fait partie d'une large gamme de travaux se référant aux centres d'appels. (Koole and Mandelbaum, 2002) ont présenté un état de l'art très complet sur les centres d'appels et les outils mathématiques servant à les modéliser. (Boxma and de Waal, 1993) ainsi que (Brandt et *al.*, 1997) ont, eux, abordé le problème des abandons des clients en attente en affectant un coût à chaque appel abandonné. (Saltzman and Mehrotra, 2001) ont montré l'efficacité de la simulation lors de la modélisation et de l'analyse des centres d'appels. (Mandelbaum et *al.*, 1999) ont analysé les abandons et les rappels avec un modèle continu qui peut décrire, dans certains cas, le régime transitoire du système. (Artalejo, 1995) a utilisé une approximation pour analyser la performance du système avec rappels.

(Garnett et *al.*, 2002) ont étudié des systèmes de files d'attente avec abandons et les ont comparés avec des files d'attente M/M/C classiques. D'autres auteurs ont approché le problème des rappels pour des objectifs différents.

Le modèle que nous analysons dans cet article comprend C conseillers de clientèle et une file de capacité totale K. Nous considérons une seule classe de clients dont la patience est supposée limitée ce qui veut dire qu'ils peuvent abandonner après un temps d'attente qui suit une loi exponentielle. Nous faisons l'hypothèse que le système atteint le régime stationnaire assez vite pour pouvoir négliger l'importance du régime transitoire ce qui nous permettra d'utiliser les probabilités stationnaires des états (de la chaîne de Markov décrivant le système). Nous supposons également que les paramètres du système (taux d'arrivée des clients, temps de service, etc.) ne varient pas dans le temps. Ceci implique que l'analyse proposée ici doit être effectuée plusieurs fois dans la journée si nécessaire. La nouveauté du modèle réside dans le fait qu'il intègre à la fois les abandons et les renouvellements d'appel.

Notre premier objectif est de montrer l'importance du phénomène de renouvellement d'appels. Pour y parvenir, nous avons démontré qu'il peut y avoir autant de rappels que d'appels frais, c'est-à-dire les appels reçus qui ne sont pas des renouvellements. Notre deuxième objectif est de dimensionner un centre d'appels en maximisant le taux de prise en charge des clients. Ce taux correspond à la proportion des appels qui ont pu accéder à un conseiller de clientèle. Nous allons montrer que le fait de ne pas considérer que les appels reçus se composent de rappels en plus des nouveaux appels peut impliquer des erreurs importantes sur le dimensionnement. Ainsi, la considération d'une file d'attente du type M/M/C/K + M sans rappel ne peut pas suffire à la détermination du nombre optimal de conseillers à planifier. Ces erreurs ont une importance considérable dans un centre d'appels vu que le personnel représente entre 60 % et 70 % de son coût total.

Dans la section 2 de cet article, nous décrivons le problème de renouvellement des appels. Nous y analyserons aussi la chaîne de Markov qui modélise le système. Cette analyse aboutira au calcul des probabilités des états au régime stationnaire. À la troisième section nous passerons à l'étude du phénomène de rappel en validant, par simulation, le modèle de chaîne de Markov étudié et en montrant que les rappels peuvent être aussi importants que les arrivées des nouveaux clients. À la section 4 nous décrivons le dimensionnement du centre d'appels en nous basant sur la chaîne de Markov déjà étudiée. Nous y comparons aussi deux systèmes, le premier fait le dimensionnement selon un critère de service en prenant en compte les rappels et l'autre ne les considèrent pas. L'article se termine par nos conclusions et nos perspectives pour le futur.

2. LE PROBLÈME DE RENOUVELLEMENT DES APPELS

2.1. Description du problème

Considérons un centre d'appels téléphoniques disposant de *C* conseillers de clientèle formant une seule classe de clients. On suppose que l'arrivée des clients est poissonienne et que les temps de service des conseillers, qui se composent d'un temps de conversation et d'un temps de traitement supplémentaire pour la clôture du dossier, suivent des lois exponentielles indépendantes. Ce centre d'appels peut donc être modélisé par une file d'attente M/M/C de taux d'arrivée λ et de taux de service μ par conseiller de clientèle. En pratique, les clients qui veulent accéder au centre d'appels peuvent abandonner en mettant fin à leur appel par leur propre initiative après avoir attendu relativement longtemps. Afin d'éviter de très longues attentes, certains centres d'appels limitent la taille de leur file d'attente. Ainsi, s'il y a déjà beaucoup de clients en attente lors de l'arrivée d'un nouveau client, alors celui-ci va être déconnecté immédiatement au lieu d'attendre le service de toutes les personnes venues avant lui. Le système ainsi modélisé peut être représenté par une file M/M/C/K + M où *K* désigne la taille totale de la file (*K*-*C* serait donc la capacité de la file d'attente) et le M supplémentaire désignant la loi markovienne qui représente les abandons des clients de la file. On suppose qu'un client abandonne la file d'attente si le service ne commence pas avant une durée suivant une loi exponentielle de taux θ que nous appèlerons par la suite taux d'abandon.

Dans cette modélisation intégrant abandons et rappels, les clients non satisfaits auront toujours la possibilité de renouveler leurs appels plus tard. Ceci va affecter les appels qui arrivent au système (appels observés) qui seront, désormais, composés de nouveaux appels (ou appels "frais") et de rappels. Nous supposons que les clients ayant abandonné ne veulent pas rappeler par la suite alors que les clients déconnectés vont le faire avec une probabilité *p* (que nous appèlerons probabilité de rappel) et ce, après un délai exponentiel de taux δ. La Figure 1 illustre le fonctionnement du centre d'appels comme nous venons de la décrire.

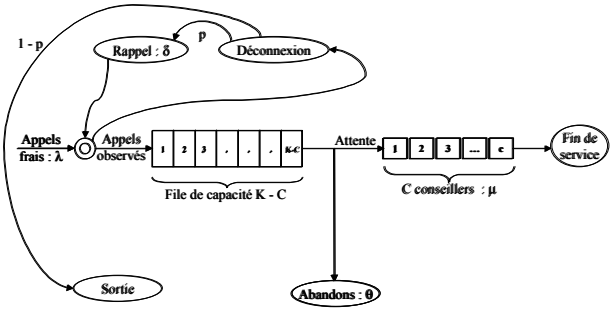

Figure 1. Schéma d'un centre d'appels avec abandons, déconnexions et rappels

2.2. Modélisation sous forme de Chaîne de Markov à Temps Continu (CMTC)

Le problème de renouvellement des appels peut être modélisé sous forme de chaîne de Markov à temps continu. Dans cette modélisation (voir Figure 2), nous avons représenté l'ensemble des états en deux dimensions. La première dimension correspond à la file réelle (qui se compose des C conseillers et de la file d'attente) où le nombre de clients présents ne peut dépasser K_1, capacité de la file. La deuxième dimension correspond, elle, aux clients déconnectés et décidés à rappeler plus tard. Cette dimension est notée « file fictive » ou encore "orbite". Ainsi, l'état (m,n), m=0,1,2,…,K_1, n=0,1,2,…,K_2, implique l'existence de m clients dans la file réelle et de n clients en orbite et qui vont rappeler plus tard d'une façon exponentielle de taux δ.

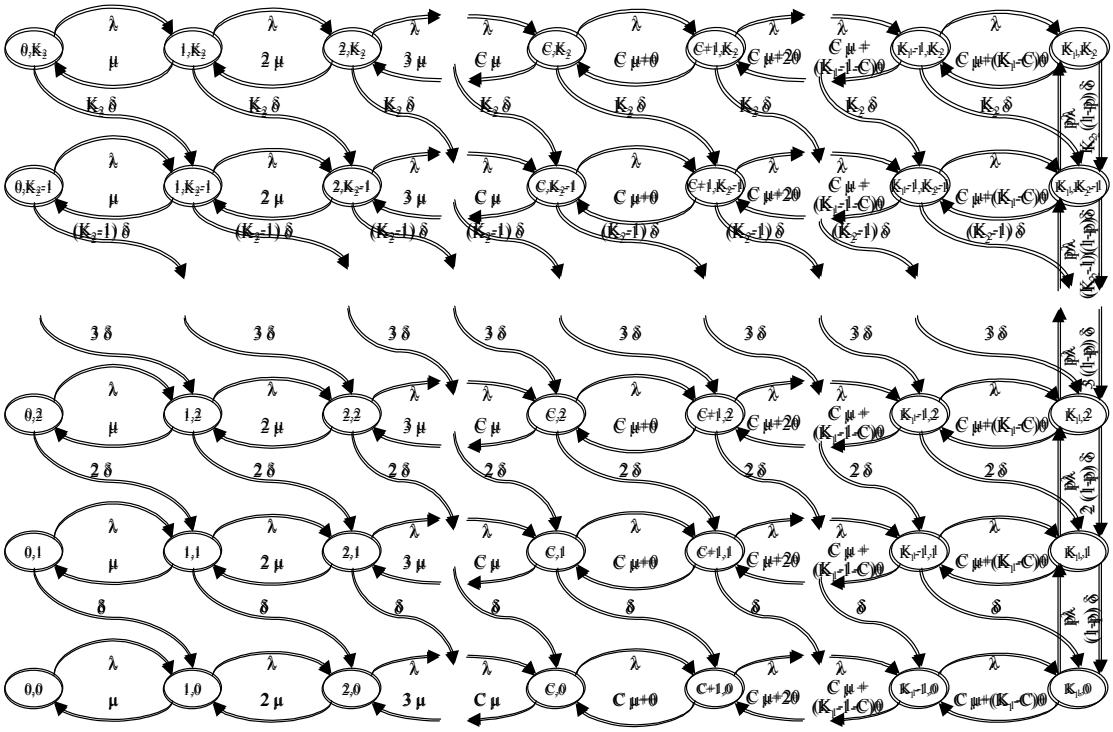

Figure 2. Modélisation du problème sous forme d'une CMTC

La chaîne de Markov obtenue ci-dessus est toujours stable, même si la charge du système ($\rho=\lambda/C.\mu$) est supérieure à 1. Cette stabilité vient du fait que plus il y a des clients en orbite plus le nombre de clients qui abandonnent après déconnexion est grand. A noter aussi que la chaîne est finie dans la dimension de la file réelle et infinie dans celle de l'orbite.

2.3. Analyse de la chaîne de Markov en régime stationnaire

Vu la dimension infinie de l'orbite, la chaîne de Markov obtenue est tronquée à une taille égale à K_2 prise suffisamment grande pour pouvoir négliger les états correspondants à une orbite supérieure. Cette CMTC obéit aux taux de transition $P_{(m,n)(i,j)}$ relatifs aux passages des états (m,n) aux états (i,j). Les taux de transition non nuls sont représentés par les formules suivantes :

– Si n=0:
$$\begin{cases} P_{(m,n)(m+1,n)}=\lambda & \text{si } m<K_1 \\ P_{(m,n)(m-1,n)}=m.\mu & \text{si } 0<m\le C \\ P_{(m,n)(m-1,n)}=C.\mu+(m-C)\theta & \text{si } C<m\le K_1 \\ P_{(m,n)(m,n+1)}=p.\lambda & \text{si } m=K_1 \end{cases} \quad (1)$$

– Si $0<n\le K_2$:
$$\begin{cases} P_{(m,n)(m+1,n)}=\lambda & \text{si } m<K_1 \\ P_{(m,n)(m+1,n-1)}=n.\delta & \text{si } m<K_1 \\ P_{(m,n)(m-1,n)}=m.\mu & \text{si } 0<m\le C \\ P_{(m,n)(m-1,n)}=C.\mu+(m-C)\theta & \text{si } C<m\le K_1 \\ P_{(m,n)(m,n+1)}=p.\lambda & \text{si } m=K_1 \\ P_{(m,n)(m,n-1)}=n.(1-p).\delta & \text{si } m=K_1 \end{cases} \quad (2)$$

– Si $n=K_2$:
$$\begin{cases} P_{(m,n)(m+1,n)}=\lambda & \text{si } m<K_1 \\ P_{(m,n)(m+1,n-1)}=n.\delta & \text{si } m<K_1 \\ P_{(m,n)(m-1,n)}=m.\mu & \text{si } 0<m\le C \\ P_{(m,n)(m-1,n)}=C.\mu+(m-C)\theta & \text{si } C<m\le K_1 \\ P_{(m,n)(m,n-1)}=n.(1-p).\delta & \text{si } m=K_1 \end{cases} \quad (3)$$

Le calcul des probabilités stationnaires $\Pi_{m,n}$ des états (m,n) peut être effectué par plusieurs méthodes. Nous avons utilisé la méthode décrite par (Tran-Gia and Mandjes, 1997), elle se compose d'un algorithme récursif. Dans cet algorithme, nous calculons, d'abord, les probabilités des états pour $n = K_2$ en commençant par $m = 1$ jusqu'à $m = K_1$ et ce, en fonction de la probabilité de l'état $(0,K_2)$. Ensuite, nous passons à la ligne suivante $(m = m + 1)$ toujours en calculant les probabilités des

états en fonction de celle de l'état $(0,K_2)$. A la fin, la condition de normalisation $\sum_{m=0}^{K_1}\sum_{n=0}^{K_2}\Pi_{m,n}=1$ nous permettra de déterminer toutes les probabilités des états au régime stationnaire.

3. ÉTUDE NUMÉRIQUE DU PHÉNOMÈNE DE RAPPEL

Dans cette section, nous validons, par simulation, la troncation de la CMTC faite dans le but d'avoir une bonne approximation du système. Nous montrons par la suite l'importance que peut avoir le phénomène de rappel à l'aide du même exemple utilisé pour la validation.

Afin de valider le modèle stochastique introduit précédemment, une série de simulations a été effectuée. La Figure 3 montre l'évolution du taux de rappel, défini par $\sum_{n=1}^{K_2} n.\delta.\sum_{m=0}^{K_1}\Pi_{m,n}$ et représentant le nombre moyen de renouvellement d'appel par unité de temps par rapport au taux d'appel frais. Nous y avons étudié deux centres d'appels semblables. L'un (100 conseillers) est quatre fois plus grand que l'autre (25 conseillers). Dans cette étude, nous avons fait varier le taux d'appel frais λ pour couvrir une charge allant de 70 % à 130 %. Nous constatons que les courbes obtenues avec le modèle stochastique coïncident avec celles que donne la simulation d'où la fiabilité des calculs effectués et de la troncation effectuée.

Ces mêmes courbes nous montrent que pour des charges supérieures à 1 nous obtenons des taux de rappel très proches du taux d'appel frais. Ainsi, pour le premier exemple, le taux de rappel est de 31 pour un taux d'appel frais égal à 39 ce qui fait un taux d'appel observé de 70. Plus tard, lors du dimensionnement du centre d'appels, le fait d'ignorer le phénomène de rappel induira des erreurs importantes. En effet, si les clients arrivent avec un taux d'appel égal à 70 alors il faudra mettre beaucoup plus de conseillers qu'il ne le faut pour satisfaire cette demande puisque la demande réelle arrive avec un taux de 39 appels par unité de temps. Ces erreurs impliqueront donc un surcoût dans le fonctionnement du centre d'appels.

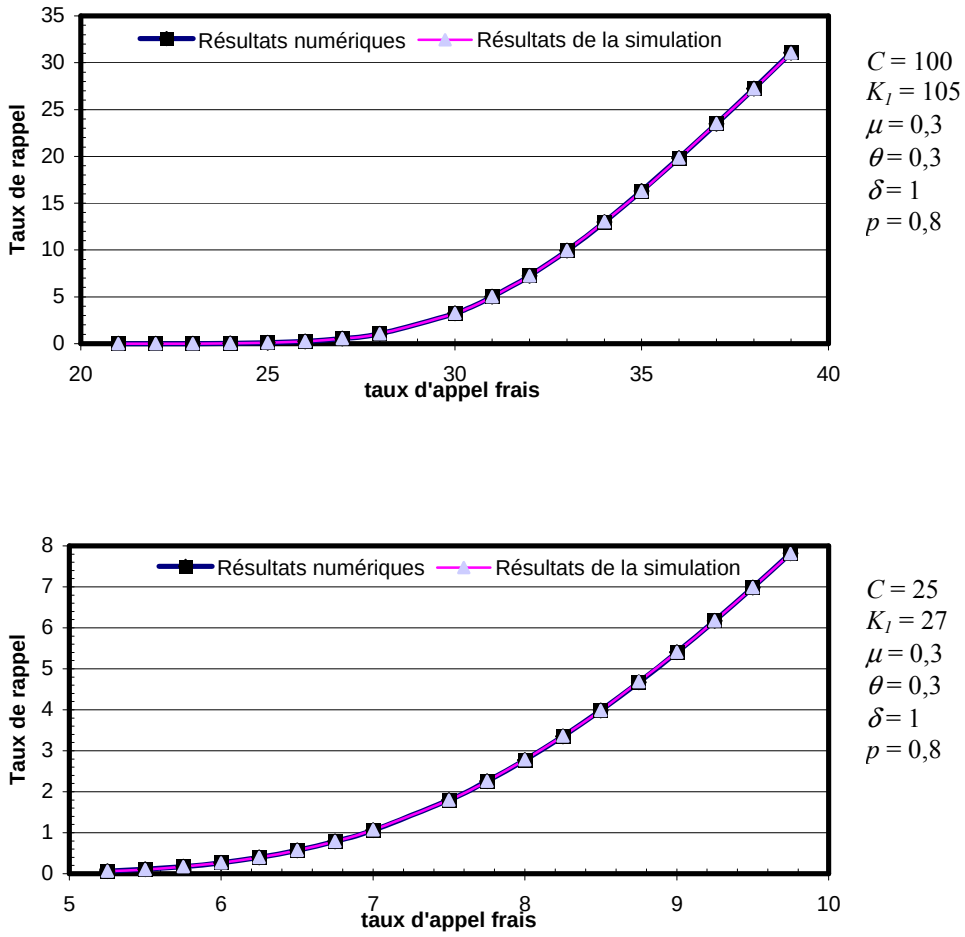

Figure 3. Évolution du taux de rappel en fonction du taux d'appel frais

4. DIMENSIONNEMENT D'UN CENTRE D'APPELS

Dans cette section, nous allons traiter l'exploitation de l'information supplémentaire sur le taux de rappel dans le but de bien dimensionner le centre d'appels. Pour cela nous fixons un critère de qualité de service à satisfaire. Cela va être un taux de prise en charge des appels, c'est à dire la proportion des appels qui ont pu rejoindre un conseiller. Cette proportion correspond aux appels non déconnectés et non abandonnés. Une comparaison des résultats obtenus avec un système de file d'attente M/M/C/K + M sans rappel est effectuée afin de voir l'intérêt de considérer les rappels dans le dimensionnement. Les données nécessaires pour effectuer la comparaison sont prises dans la base de données de l'historique sur laquelle se base, habituellement, les prévisions. Nous supposons donc que le centre d'appels dispose du taux d'arrivées enregistré sur une journée comparable de point de vue charge de travail ainsi que du nombre de conseillers présents ce jour-là et des autres paramètres déjà rencontrés.

Sans considération des rappels, nous pouvons écrire le taux de prise en charge sous la forme suivante :

$$\text{Prise en charge } (\text{sans rappels}) = 1 - \frac{(\lambda.\Pi_K + \theta.Q_W)}{\lambda} \quad (4)$$

Π_K désignant la probabilité stationnaire de trouver K clients dans le système et Q_W étant la taille moyenne de la file d'attente.

– Pour ce système sans rappels, le calcul numérique des probabilités stationnaires n'est pas très compliqué puisque c'est un processus de naissance et de mort. Cela a été analysé par (Garnett et *al.*, 2002). Suite à ce calcul, et en fonction du nombre de conseillers que nous fixons, nous pouvons déterminer le taux de prise en charge des appels. Une procédure récursive inverse permet de déduire le nombre de conseillers nécessaires pour atteindre l'objectif de qualité de service. Ceci va être comparé avec ce que l'on obtient après la considération des rappels et ce pour quatre différentes qualités de service.

– En considérant l'existence des rappels, nous savons que le taux d'appel enregistré dans l'historique est relatif aux appels observés qui se composent d'appels frais et de renouvellements d'appels. Pour dimensionner le système, nous devons commencer par déterminer le taux d'appel frais étant donné le taux d'appel observé ainsi que le nombre de conseillers présents le jour de l'extraction des données. Les autres paramètres sont supposés connus par des analyses antérieures de l'historique. Nous calculons le taux d'appel frais à l'aide d'une procédure numérique récursive que nous avons appliquée à plusieurs exemples (dont le résultat est affiché par le Tableau 1). La différence entre ces exemples se situe au niveau du nombre de conseillers se trouvant le jour où les données ont été enregistrées. Si par exemple 25 conseillers travaillaient le jour de la récolte des données alors le taux d'appels frais serait de 9,79 pour un taux d'appel observé de 15.

C	λ
25	9,79
30	10,81
35	11,78
40	12,69
45	13,48
50	14,15
55	14,62

$K = C + 5$

$\lambda_{observé} = 15$

$\mu = 0{,}3$

$\theta = 0{,}3$

$\delta = 1$

$p = 0{,}8$

Tableau 1. Évolution de λ en fonction de C

Une fois le taux d'appel frais déterminé, la prise en charge des appels est déterminée par la formule suivante :

$$\text{Prise en charge } (\text{avec rappels}) = 1 - \frac{\left(\lambda_{obs}.\sum_{n=0}^{K_2}\Pi_{K_1,n} + \theta.Q_W\right)}{\lambda_{obs}} \quad (5)$$

La Figure 4 montre l'effet de la négligence du phénomène de rappel sur le dimensionnement du système. Nous y avons tracé l'évolution du nombre de conseillers optimal en fonction de la qualité de service objectif et ce, pour trois systèmes différents. Le premier système correspond au système qui n'intègre pas la notion de rappel : le nombre de conseillers nécessaires est alors directement dépendant du taux d'arrivée d'appels observés prévus (supposés identiques à celui des appels frais prévus). Le deuxième représente ce que nous obtenons en intégrant les rappels et en partant de l'hypothèse d'un nombre de conseillers égal à 25 dans l'historique (Tableau 1) : le nombre de conseillers nécessaires dépend alors du taux d'arrivée d'appels frais prévus déduits du taux d'appels observés et des 25 conseillers présents pour y répondre. Le troisième système est identique au deuxième à part le nombre de conseillers qui passe de 25 à 55.

Dans la Figure 4, nous observons une nette différence entre les systèmes étudiés. Si par exemple nous fixions comme objectif un taux de charge supérieur à 90 %, alors nous pouvons passer de 48 conseillers (sans rappel) à 34 seulement (ou aussi à 51, cela dépend de ce que donne l'historique). Cela implique parfois un surdimensionnement, et parfois un sous dimensionnement.

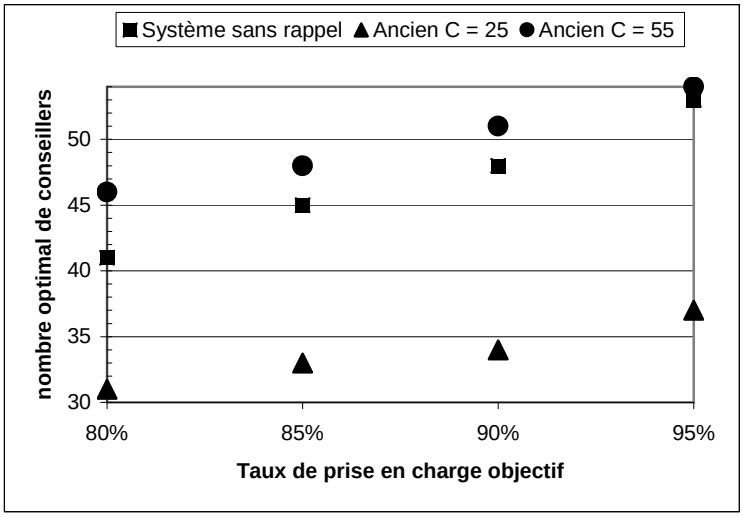

Figure 4. Effet de la considération des rappels lors du dimensionnement

5. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons mis l'accent sur la nécessité de considérer le phénomène de rappel lors du dimensionnement d'un centre d'appels. Nous avons proposé pour analyser cette problématique un modèle stochastique de chaîne de Markov à deux dimensions qui fonctionne bien mais qui devient de plus en plus délicat à utiliser pour des systèmes importants (à cause de l'augmentation du nombre d'états, ce qui nécessite un temps de calcul plus important). Nous avons en premier lieu illustré l'importance des rappels en montrant que le taux de rappel peut être de taille comparable au taux d'appel frais qui l'a engendré. Par la suite, nous avons montré les erreurs que la non prise en compte des rappels peut induire lors du dimensionnement du centre d'appels. Et comme le montre bien la Figure 4, cela peut provoquer ou bien un surdimensionnement ou bien un sous dimensionnement par rapport au niveau de service cible du centre d'appels. Ces erreurs conduisent à des objectifs non atteints ou atteints largement ce qui signifie des coûts de fonctionnement du centre plus importants que nécessaire. Pour rattraper les erreurs observées, les responsables peuvent toujours faire augmenter ou diminuer le nombre de conseillers ce qui contribue à l'ajout d'erreurs supplémentaires.

Le travail effectué dans cet article exploite des paramètres constants dans le temps. En réalité, tous les paramètres sont variables : le taux d'arrivée varie au cours d'une même journée, les conseillers de clientèle sont programmés par tranches horaires, etc. Il est intéressant d'étudier le phénomène de rappel en multi-période et, dans ce cas, il faut utiliser des modèles qui tiennent en compte l'enchaînement des périodes de la journée ainsi que le régime transitoire du système qui devient alors non négligeable par rapport au régime stationnaire.

RÉFÉRENCES

Artalejo J.R. A queueing system with returning customers and waiting line. *Operations Research Letters*, 17, pp. 191-199, 1995.

Boxma O.J. and de Waal P.R. Multiserver Queues with impatient customers. Report BS-R9319, ISSN 0924-0659, CWI, 1993.

Brandt A., Brandt M., Spahl G. and Weber D. Modelling and optimisation of call distribution systems. Teletraffic Contributions for the Information Age, Proceedings of the 15th International Teletraffic Congress - ITC 15 (Washington), Elsevier, pp. 133-144. 1997.

Garnett O., Mandelbaum A. and Reiman M. Designing a call center with impatient customers. Final version, April 2002. (To appear in *MSOM*).

Gross D. and Harris C.M. Fundamentals of queueing theory. Wiley series in probability and mathematical statistics. 3rd edition, December 1997.

Koole G. and Mandelbaum A. Queueing models of call centers An introduction. Abridged version to appear in *Annals of Operation Research*. 112. 2002.

Mandelbaum A., Massey W.A., Reiman M. I. and Rider B. Time varying multiserver queues with abandonment and retrials. ITC-16, *Teletraffic Engineering in a Competitive World*, Editors P.Key and D.Smith, Elsevier, pp. 355-364, 1999.

Saltzman R.B. and Mehrotra V. A call center uses simulation to drive strategic change. *Interfaces*, vol. 31, issue 3, pp. 87-101, May-June 2001.

Tran-Gia P. and Mandjes M. Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal on selected areas in communications*, vol. 15, no. 8, pp. 1406-1414, October 1997.