Agrégation robuste de données massives à la volée : application aux compteurs électriques communicants

Benoît Grossin*, Yousra Chabchoub**

*EDF R&D, 1 avenue du Général de Gaulle 92140 Clamart
benoit.grossin@edf.fr
** ISEP, 28 Rue Notre Dame des Champs 75006 Paris
BILab, Télécom ParisTech, 46 rue Barrault 75634 Paris Cedex 13
yousra.chabchoub@isep.fr

Résumé. Dans les années à venir, plusieurs millions de compteurs électriques communicants seront déployés sur l'ensemble du territoire français. Afin d'assurer la fiabilité d'un réseau de cette envergure nous proposons une topologie de communication multi-chemins qui repose sur la duplication des données transmises. Toute exploitation des données collectées doit alors tenir compte de la présence d'éléments dupliqués. Dans cet article, nous proposons une nouvelle méthode permettant de calculer en ligne des consommations électriques agrégées (agrégation spatiale). L'idée est d'adapter l'algorithme probabiliste *Summation sketch* de Considine et al. au contexte des compteurs communicants. Cette approche a l'avantage d'être insensible à la duplication et permet de profiter de la structure massivement distribuée du réseau de communication des futurs compteurs électriques. L'expérimentation de cette méthode sur des données réelles montre qu'elle donne une bonne précision sur l'estimation des consommations agrégées. Cette approche est aussi complétée par une méthode basée sur la théorie des sondages : On obtient une meilleure réactivité de l'estimateur avec rapidement et donc sur des données significativement partielles une erreur inférieure à 2.5%.

1 Introduction

Les pannes et les pertes de paquets sont très fréquentes dans les réseaux sans fil à cause des problèmes d'interférence, de la collision entre les paquets et de la faible puissance du signal. Plusieurs techniques ont été développées et mises en oeuvre pour limiter la perte de données et garantir une certaine fiabilité dans les transmissions. On peut citer par exemple les codes correcteurs qui sont essentiellement basés sur la redondance (voir par exemple (Caire et al. (1998))), ou encore la duplication. Dans ce papier, on se limite à l'étude de la duplication. Il s'agit d'envoyer la donnée plusieurs fois afin d'augmenter ses chances d'arriver au moins une fois à la destination. Suite à cet envoi multiple, tous les cas sont envisageables : le meilleur des cas est que la destination reçoive exactement une fois le message. Le pire des cas est que le message soit perdu malgré la duplication. Enfin, le cas le plus fréquent est que le message

soit reçu plusieurs fois (si le taux de perte est assez faible). Si on souhaite garantir la réception du message il faut ajouter un mécanisme d'acquittement, ce qui n'est pas toujours possible dans le contexte des réseaux sans fil où les ressources sont très limitées. De plus, au niveau de la destination, le traitement des données doit tenir compte du fait qu'elles peuvent se répéter. Ce problème a été largement étudié dans la littérature. Plusieurs techniques dites ODI (Order and Duplicate Insensitive) ont été conçues et utilisées dans différents domaines d'application comme par exemple les réseaux de capteurs (voir Nath et al. (2008)).

Flajolet et Martin sont parmi les pionniers dans ce domaine avec leur algorithme Probabilistic counting (Flajolet et Martin (1985)) qui a pour but d'estimer le nombre d'éléments distincts dans un multi-ensemble, c'est à dire un ensemble où un élément peut se répéter. Le point fort de cet algorithme est qu'il assure une garantie statistique sur l'erreur de l'estimation et utilise une faible mémoire. Leur approche basée sur les sketches a été fortement reprise et adaptée à différents autres contextes. Considine et al. proposent par exemple dans (Considine et al. (2009)) une adaptation de Probabilistic counting pour l'estimation des fréquences des éléments ou encore de la somme des éléments distincts dans un ensemble, dans le cadre des réseaux de capteurs.

Dans ce papier, nous avons d'abord appliqué l'algorithme summation sketch de Considine et al. à un nouveau contexte, qui est l'agrégation en temps réel des consommations électriques à une grande l'échelle. Notre but est d'estimer en ligne la consommation électrique agrégée à différentes échelles en France. Une telle connaissance est d'une grande utilité pour le pilotage du réseau et de la production électriques. Cette application est totalement innovante car elle est basée sur les *compteurs électriques communicants* qui sont encore en phase de déploiement en France. Nous avons aussi proposé une amélioration de summation sketch en introduisant un estimateur plus réactif basé sur la théorie des sondages.

Le reste du papier est organisé comme suit : Dans la section 2 nous présentons le contexte applicatif à savoir l'architecture du réseau de collecte d'informations de consommations électriques et les enjeux de cette application. Une description des algorithmes de Flajolet et al. (FM sketch) et de Considine et al. (Summation sketch) est donnée dans section 3. Dans la section 4, nous expliquons d'abord l'adaptation de l'algorithme de Considine et al. à notre contexte puis la nouvelle méthode proposée. Les résultats expérimentaux des algorithmes cités précédemment sont présentés dans la partie 5. La conclusion et les perspectives sont données dans la partie 6.

2 Contexte applicatif

Au niveau mondial, de plus en plus d'énergéticiens s'engagent actuellement dans des projets de compteurs électriques communicants, plus communément connus sous le terme anglo-saxon de smart meters : on parle alors de Smart Metering. La motivation et les enjeux de ces projets ne sont pas identiques, même si la plupart sont centrés sur des besoins de facturation, d'exploitation du réseau de distribution et de maîtrise de l'énergie. Ces nombreuses initiatives font apparaître des choix variés en termes de périmètre, de couverture fonctionnelle et de technologies utilisées.

2.1 Le projet français de compteurs communicants

Le gestionnaire du réseau de distribution d'électricité en France a initié un projet de compteurs communicants. Actuellement en phase pilote, ce projet vise à terme le déploiement de 35 millions de compteurs évolués sur le territoire français (voir (le décret 2010-1022 du 31 août 2010)). L'introduction des compteurs communicants évolués représenterait un vraie révolution pour les acteurs français de l'électricité : tout d'abord de par la diversité des données transmises (données relatives à la conduite du réseau, alarmes, données de gestion relatives à la fourniture de l'électricité, informations relatives à la consommation de chaque client), mais aussi de par leur abondance et leur fréquence de relève et de mise à disposition. L'accès, en temps réel, à ces courbes de consommation individuelle détaillée, appelées courbes de charge, laisse imaginer de nombreuses possibilités, comme par exemple, une offre de service pour les clients pour les alerter lors de la détection de consommation anormale. Une vue agrégée des consommations électriques (agrégation spatiale multi-échelles, ou agrégation temporelle) est aussi très utile pour le pilotage du réseau et de la production électrique. L'intérêt suscité par ces données devrait donner naissance à un ensemble important de services que devra offrir le système d'information (SI) du gestionnaire du réseau de distribution d'électricité, aussi bien pour ses besoins internes de distributeur, que des services proposés aux autres acteurs du marché électrique français.

2.2 Proposition d'un réseau de multi-chemins

Le périmètre national de ce projet représente un défi majeur. La qualité de la communication entre les 35 millions points de mesure et 1 système d'information (SI) unique requiert une architecture fiable et robuste. En effet, les pannes et les pertes de paquets sont très fréquentes dans les réseaux, notamment sans fil à cause des problèmes d'interférence, de la collision entre les paquets et de la faible puissance du signal.
Bien que cela ne soit pas l'option retenue pour le pilote, nous pensons qu'une architecture fortement maillée serait une solution intéressante pour assurer une bonne qualité de communication dans un réseau de compteurs de cette envergure. En effet une topologie proposant plusieurs chemins entre chaque compteur et le SI permet de déjouer efficacement la défaillance d'une transmission et d'un maillon du réseau. Plusieurs copies de chaque mesure réalisée par les compteurs seront envoyées au destinataire unique, le SI central, selon des chemins différents. Le gain en fiabilité est alors conséquent, mais engendre une duplication des messages en transit sur le réseau et reçus par le destinateur. Notons que cette proposition de réseau multi-chemin (multi-pathing) n'est pas la seule solution pour assurer la fiabilité et la robustesse d'un réseau. On peut aussi envisager de s'appuyer sur des protocoles de transmission robuste. Cet aspect n'est pas étudié dans cet article, mais quelque soit l'approche envisagée, la robustesse et la fiabilité de la communication a un coût en terme de bande passante utilisée.
Comme indiqué, l'architecture hiérarchique retenue pour le pilote en cours est différente de celle proposée ici (le SI gère une grappe de concentrateurs qui eux-mêmes gèrent chacun une grappe de compteurs) et son mode d'exploitation semble moins sensible au besoin de robustesse des communications. Les données sont collectées en mode batch quotidiennement et les contraintes de temps moins pesantes (on est dans un ordre de grandeur de la journée) permettent d'envisager d'interroger à nouveau les compteurs ou concentrateurs dont on n'aurait pas reçu les mesures. Ceci est bien adapté à des services comme la facturation où l'on peut

tolérer une latence relativement élevée mais on ne peut pas douter que le besoin de services temps-réels vont se multiplier. On devra alors s'orienter vers un modèle de mise à disposition des données de type push qui générera un flux de données massif. Le besoin d'assurer la robustesse de la transmission de ce flux sera alors crucial et ne pourra pas reposer sur le même mécanisme (identification des données manquantes, puis nouvelle transmission). C'est dans ce cadre prospectif que nous proposons un réseau de communication multi-chemins et une méthode d'agrégation des données adaptée à cette topologie qui se démarque largement des choix et solutions actuels. Notons qu'on néglige dans ce cadre le problème d'intégrité des données. Autrement dit, on suppose que les données ne sont pas erronées.

2.3 Agrégation des consommations électriques individuelles à la volée

Parmi les services rendus possibles par le déploiement national de compteurs communicants, celui du calcul d'agrégat de courbes de consommations électriques individuelles à différentes échelles (nationale, régionale, ou par segment de clientèle) est l'un des plus immédiats compte tenu de sa simplicité et de son intérêt : il s'agit d'une simple somme qui permet par exemple de mieux connaître la consommation des clients, pour des études commerciales ou des besoins réseau. C'est aussi un service qui est un candidat idéal au glissement vers un besoin temps réel évoqué précédemment. Dans le cadre d'un réseau multi-chemins que nous proposons pour sécuriser la transmission des données sous forme de flux, cet agrégat doit alors prendre en compte la présence d'éléments dupliqués et les autres caractéristiques du cadre applicatif comme l'aspect massivement distribué et la forte volumétrie en jeu. Les travaux décrits dans la suite de cet article se placent dans le cadre de cette problématique : ils posent la question de la reconstitution à la volée et en quasi temps-réel d'une courbe de charge agrégée, dite synchrone dans un environnement de compteurs communicants.

Dans la suite de cet article, nous allons étudier une méthode astucieuse pour estimer ces synchrones dans le cadre d'un réseau multi-chemin, i.e. avec une redondance des valeurs à sommer.

3 Principe d'estimation d'agrégat par les méthodes sketches

3.1 Approche sketch, un choix adapté au contexte applicatif

Calculer une somme, même avec des éléments dupliqués, ne représente pas en soi un défi majeur : Bien que l'opérateur somme soit sensible à la présence d'éléments dupliqués, on peut imaginer garder une trace des éléments reçus pour ne compter qu'une seule fois chaque valeur. C'est la méthode dite naïve. Le contexte applicatif présente une forte volumétrie, avec une somme de 35 millions de points de mesures dupliqués à sommer pour chaque instant considéré. Pour autant, accomplir cette tâche avec une approche standard ne représenterait pas une difficulté insurmontable pour un SI actuel. Mais dans notre cas, il faut considérer le calcul de somme en présence de duplication dans un contexte complet d'exploitation du SI unique. Ce SI, multi-services, sera fortement sollicité pour répondre à toutes les tâches qui lui seront attribuées. Il est donc judicieux de positionner chaque opération réalisée par le SI dans un contexte de ressources informatiques limitées afin que le SI puisse réaliser pleinement l'ensemble des tâches qui lui sont attribuées. Dans cet esprit et dans le contexte particulier des réseaux de

capteurs, Considine et al. ont proposé une approche alternative (appelée Summation sketch) reposant sur la transformation des données brutes en un objet particulier appelé sketch et dont les propriétés permettent d'estimer la somme des valeurs brutes originelles. Cette estimation est de plus parfaitement insensible à la duplication. Les ressources informatiques embarquées par les capteurs étant limitées, cette approche est économe en calcul et surtout en mémoire utilisée. Cet aspect est bien souligné par Aggarwal et Yu (Aggarwal et Yu (2007)) qui considèrent que les méthodes reposant sur des sketches sont utiles uniquement quand on a des contraintes matérielles. De plus, l'utilisation des sketches permet de distribuer la complexité du calcul d'une somme en présence d'éléments dupliqués : chaque compteur pourrait générer les sketches relatifs aux mesures qu'il fait, le SI se chargeant de les collecter dans un espace mémoire limité servant à l'estimation du résultat final. A l'opposé, l'approche naïve concentre elle l'ensemble de la complexité sur un seul point, le SI et ne profite donc pas de la structure fortement distribuée du Smart Metering. On retrouve dans l'approche sketch des éléments du paradigme " MapReduce " (Dean et Ghemawat (2008)) : les Mapper (compteurs) calculant pour chaque mesure un sketch selon un algorithme particulier, le Reducer (le SI) collectant les sketches et les agrégeant en mémoire.

Dans les paragraphes précédents, les éléments de contexte ont été posés et le choix pour une approche de type sketch justifiée. On s'intéresse maintenant aux principes de la méthode retenue. La méthode utilisée est celle proposée par Considine et al., méthode qui est une extension astucieuse des travaux de Flajolet et Martin (FM). Il convient d'ailleurs de présenter dans un premier temps ces travaux originels pour ensuite expliquer les modifications proposées par Considine et al.

3.2 Opérations de base sur des sketches

En reprenant le système de description introduit par Nath et al. dans (Nath et al. (2008)), nous utiliserons les 3 fonctions suivantes, applicables aux sketches, pour identifier les opérations de base autour des sketches.

1. $SG()$: processus de Génération d'un Sketch
2. $SF()$: retourne le résultat de la Fusion de Sketches
3. $SE()$: transforme un Sketch en Estimation du nombre d'éléments distincts

Nous allons expliciter ces fonctions dans le cadre de sketches permettant d'estimer une cardinalité puis une somme.

3.3 FM sketch pour estimer une cardinalité

Dans les années 80, Flajolet et Martin ont proposé une méthode pour estimer le nombre d'éléments distincts (cardinalité ou opérateur COUNT DISTINCT) dans un multi-ensemble de grande dimension. L'algorithme probabiliste proposé opère en une seule passe, avec peu d'opérations et se révèle économe en place mémoire nécessaire. Le calcul de cardinalité proposé par Flajolet et Martin repose sur un objet de base, le sketch (ou FM sketch) qui est un vecteur de bits. Le principe du FM sketch est de parcourir un à un les éléments e_i de l'ensemble d'éléments E dont on souhaite estimer la cardinalité $||E||$. Pour chaque e_i , on applique $SG()$ qui est la fonction de génération du sketch, qui transforme e_i en un vecteur V_i de bits. La génération de V_i se fait comme suit :

- Initialement tous les bits du vecteur V_i sont nuls

- Pour un élément e_i de E, on calcule $h(e_i)$ qui est une transformation aléatoire de e_i via une fonction de hachage h
- On note x_i, la position du 1 le plus à gauche dans la représentation binaire de $h(e_i)$
- On force $V_i[x_i]$ à 1 ($V_i[x_i] = 1$)

Notons que le processus de construction de V_i doit générer de l'aléatoire, mais doit être déterministe : en effet, deux éléments identiques doivent au travers de $SG()$ aboutir au même résultat. C'est la condition nécessaire pour assurer l'insensibilité de l'estimateur à la présence d'éléments dupliqués.
Tous les vecteurs V_i générés pour les différents éléments e_i de E seront ensuite fusionnés en un vecteur unique B avec un simple OR logique. Cette opération se fait par appel à la méthode $SF(SG(e_i), B)$ pour chaque élément e_i de E. On note alors R la position du premier bit nul dans le vecteur B obtenu suite aux opérations de fusion.

Pour avoir une estimation robuste, (Flajolet et Martin, 1985) utilisent le concept de moyenne stochastique (Stochastic Averaging). R est alors calculé sur m vecteurs B indépendants et que l'on note $B_{[1..m]}$. On obtient ainsi m valeurs de R notées $R_{[1..m]}$. L'estimation de la cardinalité de E est enfin donnée par la formule suivante :

$$SE(B) = m\frac{1}{\phi}2^{1/m\sum_{i=1}^{m}R_m}$$

Où $\phi \approx 0.77$. L'erreur standard de cet estimateur est approximativement $0.78/\sqrt{m}$. Notons que cet indice de précision est intéressant comparativement à ceux de méthodes alternatives (Durand et Flajolet (2003), Flajolet et al. (2007), Fusy et Giroire (2007)).

3.4 Summation sketch pour estimer une somme

Dans (Considine et al. (2009)), les co-auteurs se donnent la problématique suivante : on dispose de N sources de données. A chaque instant t, chaque source i envoie une valeur entière notée $val_i(t)$. On cherche à calculer la somme des N valeurs ($\sum_{i=1}^{N} val_i(t)$) pour tracer la courbe de consommation agrégée en fonction du temps. On fait ici l'hypothèse que toutes les sources sont bien synchrones. On est aussi dans un contexte où les données peuvent être dupliquées. Considine et al. proposent d'adapter la méthode FM sketch de Flajolet et al. à cette problématique. Les modifications portent uniquement sur le processus de génération des sketches $V = SG()$, le reste de l'algorithme reste inchangé. Le traitement étant le même à chaque instant t, on se passera de la notion du temps dans ce qui suit pour alléger les notations. Pour chaque source i, la génération de $V_i = SG < i, val_i >$ revient à insérer val_i éléments distincts dans le sketch V_i. Autrement dit V_i est le résultat de l'opération de fusion suivante : $SF(SG < i, e_1 >, SG < i, e_2 >, ...SG < i, e_{val_i} >)$. Ensuite les N sketches V_i seront fusionnés pour obtenir l'équivalent du vecteur B. Le déroulement du reste de l'algorithme sera le même que dans FM sketch : l'estimation de la somme sera donnée à partir du paramètre R calculé sur le vecteur B. Notons qu'ici on applique $SG()$ sur le couple $< i, val_i >$ car si deux sources i et j ont la même valeur val, il faut la compter deux fois dans la somme. Un élément dupliqué est un couple $< i, val_i >$ qui se répète. Le processus de génération de V_i, ainsi défini, nécessite val_i appels à la fonction $SG()$ ce qui est très coûteux pour les valeurs élevées de val_i. Pour remédier à ce problème Considine et al. proposent une stratégie de simulation du résultat obtenu par l'insertion de val_i éléments distincts en se basant sur la loi géométrique

(voir détails dans Considine et al. (2009)). L'impact de cette simplification est négligeable sur la valeur prise par R et l'estimateur $SE()$ reste asymptotiquement sans biais.

4 Sketches et compteurs communicants : méthode standard et proposition d'une nouvelle méthode

4.1 Positionnement et application des sketches dans l'architecture des compteurs communicants

Replaçons les Summation sketches, dont le fonctionnement et les caractéristiques ont été explicités dans la section précédente, dans le contexte du Smart Metering. Chaque mesure réalisée par un compteur est traitée par la fonction $SG()$ afin de générer un sketch. Ce sketch ne se substitue pas à la mesure brute envoyée par le compteur au SI mais vient la compléter. Ainsi basiquement, les messages envoyés par les compteurs contiennent un identifiant du compteur, une mesure brute de consommation électrique horodatée, et le sketch associé. Un schéma synthétique de l'architecture du réseau de collecte est proposé dans la figure 1. Le SI collecte les mesures brutes d'une part et traite les sketches dans un espace mémoire limité d'autre part (m*taille d'un sketch, par pas de temps distinct) afin de proposer le service d'estimation de consommations électriques agrégées en quasi temps-réel. Entre les sources de données (les compteurs) et le destinataire (le SI) la topologie de réseau multi-chemin assure la fiabilité de la communication et, comme vu précédemment, l'estimation de la somme basée sur les Summation sketches n'est pas perturbée par la présence d'éléments dupliqués.

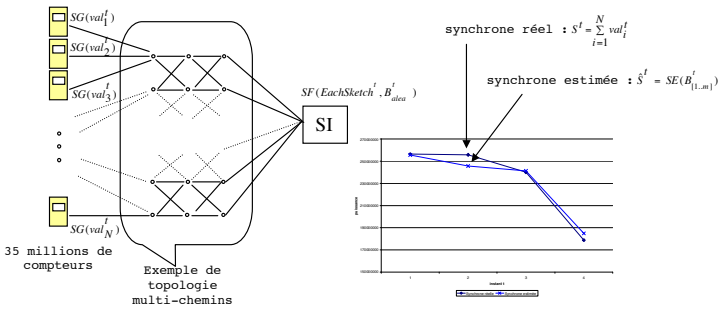

FIG. 1 – *Schéma de l'architecture proposée, positionnement des fonctions sketch et principe d'estimation de la synchrone*

Cette configuration qui est une transposition des travaux de Considine au domaine des compteurs communicants sera testée dans la section expérimentation afin de quantifier les performances d'estimation obtenues, notamment en discutant de l'impact de la mémoire allouée au niveau des SI pour agréger les sketches reçus en masse.

4.2 Vers une méthode plus réactive : proposition d'un nouvel estimateur

Dans le but de proposer un service d'estimation des consommations agrégées en quasi temps-réel, il est frustrant d'avoir à attendre la réception au niveau SI de l'ensemble des copies des mesures d'un instant t pour connaître l'estimation de la somme des mesures.

L'idée que nous proposons est d'utiliser un estimateur très courant en théorie des sondages en s'appuyant sur les sketches. Il s'agit de l'estimateur d'Horvitz-Thompson (HT), qui estime une somme d'une population à partir d'une sous-population aléatoire. Nous nous plaçons dans le cadre des plans de sondage aléatoire simple à probabilité d'inclusion égale. Sous l'hypothèse que les copies de mesures d'un instant t arrivent aléatoirement au niveau du SI, on dispose à tout moment d'un échantillon aléatoire des mesures des compteurs à un instant t, cet échantillon passant au fur et à mesure, d'une seule mesure (lors de la réception de la première mesure relative à un instant t) à l'ensemble des mesures de la totalité des compteurs (souvent en plusieurs exemplaires compte tenu de la structure multi-chemin). On a donc, à tout moment et pour l'instant t considéré :

- La somme de la sous-population des compteurs dont on a reçu au moins une copie de la mesure réalisée, via le mécanisme de Summation sketch.
- La taille de l'ensemble de tous les compteurs, que l'on suppose fixe.

Si l'on se réfère à la formule de l'estimateur HT, il manque la taille de la sous-population, c'est-à-dire le nombre de compteurs pour lesquels au moins un message de mesure a été réceptionné. Cette valeur peut être accessible en utilisant le FM sketches, qui estime le nombre d'éléments distincts (i.e. de compteurs).
Nous proposons alors l'estimateur HT modifié suivant, reposant à la fois sur les FM et Summation sketches avec :

$$\overbrace{S_{Sample} * \frac{N}{n_{Sample}}}^{Estimateur\ HT\ original} \Longrightarrow \overbrace{\hat{S}_{SummationSketches} * \frac{N}{\hat{n}_{FMSketches}}}^{Estimateur\ HT\ modifié}$$

- N, la taille totale de la population totale considérée, soit ici le nombre total de compteurs
- n_{Sample}, la taille de la population échantillonnée
- S_{Sample} , la somme de la variable considérée sur cette population échantillonnée
- $\hat{n}_{FMSketches}$, le nombre de compteurs estimé par FM sketches dont on a réceptionné au moins une mesure pour l'instant t considéré
- $\hat{S}_{SummationSketches}$, la somme estimée par les Summation sketches des mesures reçues

Notons que $\hat{n}_{FMSketches}$ et $\hat{S}_{SummationSketches}$ évoluent au fil de la réception des mesures.
Deux remarques importantes par rapport à cette proposition d'utilisation conjointe de la théorie des sondages et des sketches :

1. Le FM sketch alourdit le message envoyé par rapport à la proposition précédente et l'espace mémoire nécessaire au niveau du SI.
2. L'estimateur HT possède son propre intervalle de confiance, mais notre proposition introduit deux nouveaux éléments d'imprécision : les estimations de $\hat{n}_{FMSketches}$ et de $\hat{S}_{SummationSketches}$. Dans ces conditions et compte tenu de sa forme, l'estimateur présente une précision dégradée par rapport à l'estimateur HT original.

Cet estimateur, plus réactif, sera aussi testé dans la section relative aux expérimentations.

5 Expérimentation

Le flux de mesures utilisé pour cette expérimentation correspond à la consommation électrique de $N_c = 358,800$ clients sur une journée prise au hasard. Notons qu'il s'agit en fait de données réelles de consommation d'un ensemble limité de clients qui a servi à générer ce volume conséquent de mesures. Le pas de temps est horaire, soit 24 mesures par compteur. Les algorithmes cités ont été codés en Java via le système de gestion de flux de données StreamBase [1] . Nous n'avons pas encore simulé différentes topologies de réseau, mais les données ont bien été dupliquées (6 fois) afin de vérifier l'insensibilité des sketches à ce phénomène. La perte de données n'a pas été simulée dans l'expérimentation présentée dans cet article. Notons cependant que des expérimentations préliminaires intégraient la simulation de perte de données et ont permis de valider l'intérêt de la duplication comme solution à ce problème : en effet, le taux de réussite de transmission sont conformes aux attentes théoriques.

5.1 Performance de summation sketch dans le contexte applicatif

Dans les conditions de l'expérimentation présentées dans le paragraphe précédent, les résultats obtenus sont probants (voir figure 2).

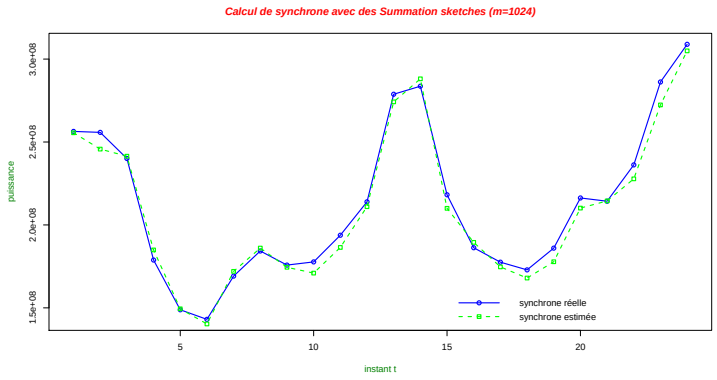

FIG. 2 – *Estimation de la courbe de consommation électrique agrégée avec les Summation sketches* ($m = 1024$)

Comme souligné précédemment, la précision de l'estimation obtenue est liée à la valeur de m, i.e. le nombre de vecteurs B que doit maintenir en mémoire le SI pour chaque pas de temps. Pour rappel, l'erreur standard de l'estimateur est $0.78/\sqrt{m}$. Dans un contexte de réseau de capteurs, la mémoire est un critère fortement limitant et la valeur de m se trouve alors naturellement bornée. Dans le contexte d'utilisation proposé dans cet article, une méthode parcimonieuse pour l'estimation de synchrone au niveau de SI est souhaitable pour ne pas concurrencer les autres services proposés, mais la contrainte d'empreinte mémoire est largement moins forte que sur un capteur. C'est pourquoi il est tentant d'utiliser des valeurs de m conséquentes afin de proposer une estimation très précise.

1. www.streambase.com

La figure 3 présente la qualité de l'estimation obtenue en fonction de m, en se basant sur l'indicateur MAPE (Mean Absolute Percent Error). Les performances des Summation sketches sont conformes aux attentes jusqu'à environ $m = 8,000$ (soit environ 50 valeurs aléatoires stockées dans chaque vecteur de $B_{[1..m]}$). Ensuite les performances se dégradent fortement. Selon nous, ce seuil correspond à la taille de l'échantillon nécessaire dans chaque vecteur de $B_{[1..m]}$ pour que les résultats soient assez significatifs. En dessous, on a un déséquilibre trop important dans les contenus des vecteurs de $B_{[1..m]}$ et le principe de la moyenne stochastique n'est plus valide. D'ailleurs Considine et al. soulignent ce phénomène et proposent un mécanisme visant à équilibrer les contenus des vecteurs de $B_{[1..m]}$. Cependant ce mécanisme alourdit la méthode (par exemple l'opération $SF()$ passe d'une complexité de $O(1)$ à $O(m)$). Dans un contexte de déploiement national, $N \approx 100N_c$: on pourrait donc utiliser des valeurs de m jusqu'à environ 800,000 soit une erreur standard théorique associée de $0,1\%$. Notons que notre proposition d'utilisation des sketches, même dans ces conditions de données massives, passerait à l'échelle puisque que ce sont les sources (et non pas le SI) qui portent la principale complexité du processus des sketches, à savoir leur génération.

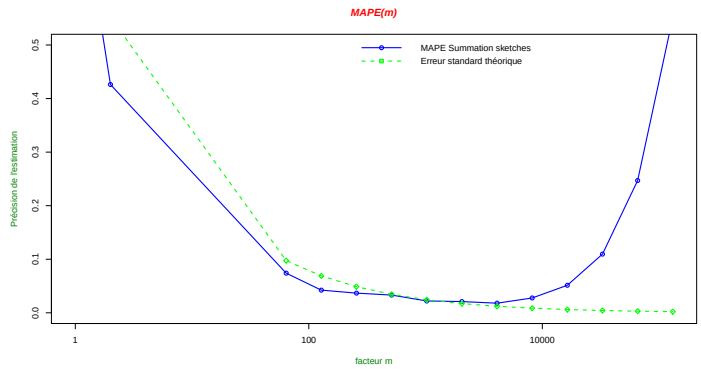

FIG. 3 – *Impact de l'espace mémoire alloué (facteur* m*) sur la qualité de l'estimation*

5.2 Performance du nouvel algorithme basé sur la théorie des sondages

La mise en application de notre proposition basée sur la théorie des sondages donne les résultats suivants (voir figure 4) pour les mesures de l'instant $t = 1$ (les mesures n'ont pas été dupliquées dans cet exemple et la valeur de m est fixée à $5,000$). La combinaison des sketches et de la théorie des sondages permettent d'estimer la synchrone au fur et à mesure de l'arrivée des mesures, et dans cet exemple on a une estimation à 2.5% de la valeur de la somme finale dès que le SI a collecté plus de la moitié des mesures relatives à l'instant t considéré.

Ainsi la nouvelle méthode proposée donne de bons résultats sur le jeu de données utilisé. Notons que la distribution des valeurs des compteurs dans les données considérées est assez régulière. Ceci peut être déduit à partir de la figure 4, où la somme sur toutes les valeurs est proportionnelle au nombre de valeurs reçues. Autrement dit, les valeurs des compteurs sont plutôt homogènes, ce qui représente un cadre adapté à l'application de la théorie des sondages. Cette propriété reste vraie même si la taille de la population augmente.

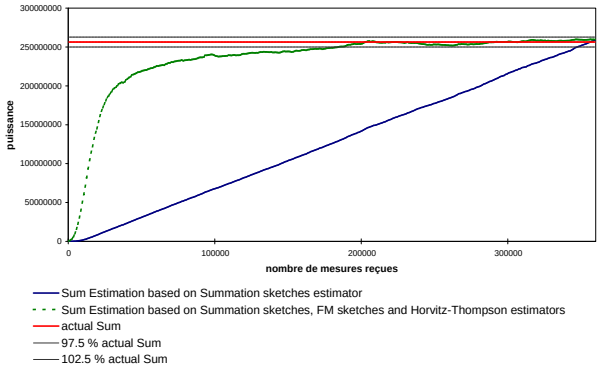

FIG. 4 – *Performance du nouvel estimateur basé sur la théorie des sondages*

6 Conclusion et perspectives

Dans cet article nous avons proposé d'utiliser une topologie multi-chemins pour assurer la fiabilité d'un réseau de grande envergure de compteurs communicants. Cette robustesse engendre la duplication des messages en transit et reçus par le SI responsable entre autres de la collecte des mesures de consommations électriques individuelles. Toute exploitation des ces données doit tenir compte de la présence de ces éléments dupliqués. Nous avons vu que pour la problématique qui nous intéresse, à savoir le calcul de synchrone au fil de l'eau (soit le calcul de la somme des consommations électriques individuelles) l'utilisation des Summation sketch représente une solution élégante et adaptée au contexte : les traitements sont massivement distribués et le SI est alors peu sollicité, à la fois en terme de calcul et d'empreinte mémoire. Ceci garantit le passage à l'échelle nécessaire dans le contexte envisagé. Summation sketch est un estimateur de l'agrégat somme parfaitement insensible à la présence d'éléments dupliqués et nous avons vu que les garanties de précision que l'on pouvait obtenir sont tout à fait acceptables bien que bornées (erreur standard pouvant être réduite jusqu'à 0.1% pour le périmètre envisagé). Afin de rendre cet estimateur plus réactif, nous avons proposé une utilisation conjointe des FM et Summation sketch et de la théorie des sondages pour estimer l'agrégat final sur des données partielles. Les premiers résultats obtenus sont encourageants mais demandent une étude plus approfondie, notamment pour quantifier les intervalles de confiance associés.
Notons enfin qu'un travail pour préciser la topologie à retenir est en cours et pose notamment la question de l'utilisation des noeuds intermédiaires du réseau : simple relais des sketches à transmettre au SI ou espaces potentiels de stockage intermédiaire de résultats partiels.

Références

Aggarwal, C. et P. Yu (2007). A survey of synopsis construction in data streams. *Data Streams*, 169–207.

Caire, G., G. Taricco, et E. Biglieri (May, 1998). Bit-interleaved coded modulation. *IEEE Transactions on information theory 44*(3).

Considine, J., M. Hadjieleftheriou, F. Li, J. Byers, et G. Kollios (2009). Robust approximate aggregation in sensor data management systems. *ACM Trans. Database Syst. 34*(1), 1–35.

Dean, J. et S. Ghemawat (2008). MapReduce : Simplified data processing on large clusters. *Communications of the ACM 51*(1), 107–113.

Durand, M. et P. Flajolet (2003). Loglog counting of large cardinalities. *Proceedings of European Symposium on Algorithms*.

Flajolet, P., E. Fusy, O. Gandouet, et F. Meunier (2007). Hyperloglog : the analysis of a near-optimal cardinality estimation algorithm. *Proceedings of the 13th conference on analysis of algorithm (AofA 07)*.

Flajolet, P. et G. Martin (1985). Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences 31*(2), 182–209.

Fusy, E. et F. Giroire (January 2007). Estmating the number of active flows in a data stream over a sliding window. *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithmics and Combinatorics (ANALC0)*, 223–231.

le décret 2010-1022 du 31 août 2010. *Calendrier de déploiement des dispositifs de comptage sur les réseaux publics d'électricité, http ://www.legifrance.gouv.fr*.

Nath, S., P. Gibbons, S. Seshan, et Z. Anderson (2008). Synopsis diffusion for robust aggregation in sensor networks. *ACM Transactions on Sensor Networks (TOSN) 4*(2), 1–40.

Summary

In the coming years, several millions of communicating electric meters will be deployed in France. To guarantee the robustness of a so huge network, we propose a multi-path topology based on the duplication of the transmitted data. Thus, the analysis of the collected data must take into account the presence of duplicated elements. In this paper, we propose a new method for an online accounting of the aggregated electric consumption (spatial aggregation). The idea is to adapt the probabilistic algorithm *Summation sketch* of Considine et al. to the context of the communicating electric meters. This approach has the advantage to be insensitive to the duplication and to allow the exploitation of the widely distributed structure of the communication network of the future communicating electric meters. The experimentation of this method on real data shows that it gives a good accuracy on the estimation of the aggregated consumptions. This approach is also extended by a method based on sampling theory: We obtain a better reactivity of the estimator with an error less than 2.5%, even on significant partial data.