Echantillonnage spatio-temporel de flux de données distribués

Raja Chiky*, Jérôme Cubillé**, Alain Dessertaine**,
Georges Hébrail*, Marie-Luce Picard **

* GET-ENST Paris
Laboratoire LTCI - UMR 5141 CNRS - Département Informatique et Réseaux
46 rue Barrault, 75634 Paris Cedex 13
Email: prenom.nom@enst.fr
** EDF R&D - Départements ICAME et OSIRIS
1, Avenue du Général de Gaulle, 92140 Clamart
Email: prenom.nom@edf.fr

Résumé. Ces dernières années, sont apparues de nombreuses applications, utilisant des données potentiellement infinies, provenant de façon continue de capteurs distribués. On retrouve ces capteurs dans des domaines aussi divers que la météorologie (établir des prévisions), le domaine militaire (surveiller des zones sensibles), l'analyse des consommations électriques (transmettre des alertes en cas de consommation anormale),... Pour faire face à la volumétrie et au taux d'arrivée des flux de données, des traitements sont effectués 'à la volée' sur les flux. En particulier, si le système n'est pas assez rapide pour traiter toutes les données d'un flux, il est possible de construire des résumés de l'information. Cette communication a pour objectif de faire un premier point sur nos travaux d'échantillonnage dans un environnement de flux de données fortement distribués. Notre approche est basée sur la théorie des sondages, l'analyse des données fonctionnelles et la gestion de flux de données. Cette approche sera illustrée par un cas réel : celui des mesures de consommations électriques.

1 Motivations

Les entrepôts de données sont de plus en plus alimentés par des flux de données provenant d'un grand nombre de capteurs distribués. Malgré l'évolution des nouvelles technologies de traitement et de stockage des données, il reste difficile voire impossible de conserver la totalité de l'information. Pour faire face à cette inflation, de nombreux travaux (Aggarwal, 2007; Babcock et al, 2002; Muthukrishnan, 2005) ont été menés ces dernières années sur la gestion et l'analyse de flux de données : un flux de données est défini comme une séquence continue, potentiellement infinie, de n-uplets (d'enregistrements) ayant tous la même structure. L'ordre d'arrivée des n-uplets n'est pas contrôlé, et les données, de par l'importance de leur volume et de leur débit d'arrivée, ne peuvent pas exhaustivement être stockées sur disque : les données passent, et doivent être traitées 'à la volée'.

Les applications actuelles des approches de traitement de flux de données portent surtout sur la supervision de systèmes, sur le déclenchement d'alarmes en temps réel, ou plus généralement sur la production de synthèses d'aide à la décision à partir de plusieurs flux. Les systèmes de gestion de flux de données (SGFD) (Stonebraker et al., 2005; STREAM; StreamBase; Golab et al., 2003) permettent de gérer facilement, et de façon générique les flux de données en inversant le rôle des données et des requêtes par rapport aux systèmes de gestion de bases de données classiques : les données sont volatiles et non prédictibles ; les requêtes sont statiques et s'exécutent en permanence sur ces données. On parle de requêtes continues. Un flux étant potentiellement infini, il est nécessaire de définir des fenêtres d'intérêt sur le flux, en particulier des fenêtres temporelles le plus souvent glissantes.

Le caractère distribué des flux, rajouté au fait qu'ils sont potentiellement infinis et arrivent avec un débit élevé, implique que de nombreux traitements nécessitent la mise en place de techniques d'approximation dont la qualité de résultat sera ajustée en fonction des ressources disponibles. On retrouve cette problématique dans le domaine des consommations électriques. En effet, avec le déploiement envisagé à moyen terme de compteurs communicants chez les clients des fournisseurs d'électricité, les consommations d'énergie électrique pourront être télérelevées à des pas de temps pouvant aller jusqu'à la seconde. Ceci permettra d'effectuer des opérations tels que la facturation, l'agrégation, le contrôle,... Il n'est cependant pas envisageable de récupérer tous ces flux à cause de la volumétrie (plus de 30 millions de compteurs) et des coûts d'exploitation.

Nous formalisons le problème posé dans le cadre de cette communication comme suit :

Nous disposons de N capteurs reliés à un système central et qui envoient une séquence de mesures temporelles en continu. Nous supposons que les données sont générées régulièrement et que les mesures sont numériques et unidimensionnelles. Chaque flux est découpé en périodes (fenêtres temporelles) de même taille. Une fenêtre temporelle est constituée de p éléments (on suppose constant le nombre d'éléments d'une fenêtre car le flux est régulier). Dans la suite, on utilise le terme « courbe » pour qualifier chaque séquence temporelle. Nous supposons que le système central a une limite de stockage par période de temps à ne pas dépasser, soit s cette limite ($s < N * p$). L'objectif de ce travail est d'approcher les courbes originales à partir d'un ensemble de données sélectionnées en respectant la limite de stockage au niveau central. Plusieurs travaux ont été menés ces dernières années sur l'échantillonnage et la construction de résumés à partir de flux de données.

2 Définitions

2.1 Echantillonnage temporel

L'échantillonnage temporel consiste à conserver habilement un nombre réduit de mesures sur une courbe, sans perdre trop d'information sur la forme de la courbe. Remarquons que si aucun échantillonnage temporel n'est réalisé, alors agréger les courbes de la population P consiste, à chaque instant t, à sommer l'ensemble des valeurs de consommation $C_i(t)$ mesurées sur chaque courbe i de la manière suivante :

$$C(t) = \sum_{i=1}^{N} C_i(t) \tag{1}$$

Le fait d'échantillonner de manière temporelle les courbes nécessite de les estimer à partir des mesures sélectionnées par interpolation. Un tel cas de figure peut être rapproché des travaux autour de l'analyse des données fonctionnelles (Deville, 1974; Ramsay et Silverman, 2005). Ces interpolations utilisent des bases fonctionnelles (comme les splines, les ondelettes, ou des transformations de fourier). Les analyses se font alors dans le nouveau système de coordonnées issu de la base de fonctions choisie.

Dans notre cas, nous proposons de remplacer dans (1) la valeur inconnue en un instant t par sa valeur interpolée :

$$\tilde{C}(t) = \sum_{i=1}^{N} (\tilde{C}_i(t) + \epsilon_i(t)) \tag{2}$$

Avec $\epsilon_i(t)$ l'erreur d'interpolation pour la courbe i à l'instant t.

2.2 Echantillonnage spatial

L'échantillonnage spatial consiste à conserver habilement un certain nombre de courbes - ce qui revient à une problématique similaire à celles abordées dans la théorie des sondages.

Dans ce cadre, nous travaillons sur la base d'une population finie (l'ensemble des N capteurs). Nous construisons une stratégie d'échantillonnage sur cette base afin d'utiliser au mieux les informations disponibles sur les capteurs pour bien estimer, sur l'échantillon qui sera finalement construit, la ou les variables d'intérêt (ici, la variable d'intérêt est la courbe globale). Citons, parmi ces stratégies, les échantillons stratifiés, les échantillons à probabilités inégales, et les échantillons équilibrés (pour plus de détails, voir (Ardilly, 2006; Tillé, 2006)).

Pour ce faire, nous avons besoin, au minimum, de la liste de tous les capteurs i, auxquels nous affectons une probabilité d'inclusion π_i dans notre échantillon. Un estimateur sans biais souvent utilisé est l'estimateur d'Horvitz-Thompson. Ainsi, nous pouvons calculer l'estimateur du total d'une variable X, dont les valeurs x_i sont mesurés seulement pour les individus i appartenant à l'échantillon que nous nommerons S de la manière suivante :

$$\hat{X}_{HT} = \sum_{i \in S} \frac{x_i}{\pi_i} \tag{3}$$

Et l'estimation de sa variance est donnée par :

$$\hat{Var}(\hat{X}_{HT}) = \sum_{i \in S} \sum_{j \in S} \left(\frac{x_i}{\pi_i} \frac{x_j}{\pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \tag{4}$$

Avec π_{ij} la probabilité d'inclusion double qu'un couple (i, j) appartienne conjointement à l'échantillon. Cette probabilité dépend de la stratégie utilisée lors de l'échantillonnage. L'estimateur de la variance présenté ci-dessus est souvent délicat à calculer de manière exacte ; des approximations par linéarisation, ou par bootstrap sont proposées dans la littérature (Deville, 1999; Ardilly, 2006).

Dans le cas d'une extrapolation de courbes sans échantillonnage temporel, il suffit de remplacer x_i et x_j par $C_i(t)$ et $C_j(t)$ pour un instant t de la fenêtre temporelle.

2.3 Exemples d'application aux flux de données distribués

Plusieurs travaux de recherche ont été menés pour développer des algorithmes de résumés des données au fil de l'eau appliqués à des flux de données distribués. Quelques travaux échantillonnent les sources de flux de données (*échantillonnage d'individus* ou *échantillonnage spatial*). Dans ce cadre, un exemple d'application est donné dans (Willett et al., 2004). En effet, une première phase active un sous-ensemble de capteurs pour récupérer leurs données. Ceci fournit une estimation de l'environnement en détectant les corrélations entre les capteurs. Puis une deuxième phase dite d'amélioration permet d'activer sélectivement des capteurs supplémentaires afin d'améliorer les estimations.

D'autres travaux échantillonnent les données provenant de chaque source (*échantillonnage temporel*). Une approche concernant l'échantillonnage temporel est présentée dans (Jain et al., 2004). Dans cet article, un filtre de Kalman est utilisé au niveau de chaque capteur pour effectuer des prédictions. Une granularité temporelle (un pas d'échantillonnage) est ajustée selon l'erreur de prédiction. Si ce pas dépasse un seuil alloué par le système central, un programme d'optimisation est lancé au niveau central afin d'affecter de nouveaux seuils aux capteurs.

3 Approche proposée et conséquences sur les estimations

3.1 Echantillonnage spatio-temporel

Dans les deux cas d'échantillonnage cités ci-dessus (échantillonnage temporel et échantillonnage spatial), on n'utilise pas les vraies valeurs résumées mais des estimations (interpolation dans le premier cas, estimations entachées d'une erreur d'échantillonnage dans le deuxième). Nous proposons une stratégie qui consiste à combiner ces deux approches, par souci d'avoir des estimateurs des courbes agrégées de qualité acceptable (avec des erreurs d'interpolation et/ou d'échantillonnage faibles). Nous nommerons cette stratégie échantillonnage spatio-temporel.

Nous formulons le problème de l'échantillonnage temporel comme suit : soit m la plus grande granularité temporelle permise pour récupérer les données. On cherche à déterminer un planning de collecte de données à effectuer pendant une période t sur les N capteurs, en effectuant un échantillonnage temporel au sein de chaque capteur. Pour cela, nous avons modélisé le problème sous forme d'un programme d'optimisation linéaire qui affectera de façon optimale une granularité temporelle inférieure ou égale à m à appliquer à chaque capteur pendant la période t, en se basant sur les caractéristiques des données de la période $t-1$. Pour plus de détail sur la modélisation et la résolution du problème, le lecteur peut se référer à (Chiky et Hébrail, 2007).

Nous proposons d'intégrer une approche sondage à notre programme d'échantillonnage temporel afin de constituer un échantillon spatio-temporel. Comme pour les approches traditionnelles, on établit un plan de sondage pour constituer un échantillon avec un nombre de capteurs n. Deux cas de figure peuvent se présenter :

- $n * p \geq s$: la taille de l'échantillon n implique le dépassement du seuil s et on est en mode dit *saturé*. Dans ce cas on applique un échantillonnage temporel optimisé afin de récupérer des courbes à granularités variables dans le temps pour chaque capteur de l'échantillon. Ceci permet de capter les périodes de forte variabilité des données. Mais, se pose un problème d'estimation des erreurs d'échantillonnage auxquelles s'ajoutent

des erreurs d'interpolation des courbes à partir des points résumés. Ce problème sera évoqué dans la section suivante.

– $n * p < s$: la taille de l'échantillon n ne dépasse pas le seuil s et on est en mode dit *non saturé*. Les courbes de l'échantillon sont prises entièrement et on applique l'échantillonnage temporel au reste des capteurs (N-n) ne faisant pas partie de l'échantillon, avec un seuil de $s' = s - n * p$. Ceci permet d'améliorer les estimations calculées à partir de l'échantillon ou d'améliorer certaines requêtes comme celles sur des sous-populations, correspondant alors à des estimateurs sur des petits domaines.

3.2 Conséquences sur les estimations

D'un point de vue formel, il suffit d'estimer une courbe globale à partir d'un échantillon S. Nous pouvons élaborer un estimateur à partir d'un estimateur de type Horvitz-Thompson :

$$\hat{\tilde{C}}_{HT}(t) = \sum_{i \in S} \frac{\tilde{C}_i(t) + \epsilon_i(t)}{\pi_i} \tag{5}$$

La principale question qui se pose à nous à ce stade est : la stratégie « spatio-temporelle » proposée précédemment peut-elle être plus performante qu'une stratégie avec échantillon simplement spatial ou simplement temporel ? D'une manière évidente, la réponse va dépendre de la stratégie d'échantillonnage spatial choisie, mais elle peut aussi dépendre des techniques de redressement habituellement utilisées en sondage pour améliorer l'estimateur (au sens de la réduction de la variance d'échantillonnage). Elle va dépendre aussi de la technique d'interpolation utilisée.

Le fait de ne pas travailler sciemment avec les vraies mesures, mais avec des mesures entachées d'une erreur aléatoire, que nous supposerons maitrisée, met à mal le cadre général d'application de l'estimateur d'Horvitz Thompson, en espérance et en variance. Le caractère non biaisé, et la formule d'estimation de sa variance présentée en (4) ne sont valables que si les mesures effectuées sur l'échantillon ne sont pas aléatoires, ce qui n'est pas le cas ici de par la présence de $\epsilon_i(t)$. Il nous faut donc connaître les performances de l'estimateur (5). En effet, il a été prouvé (Dessertaine, 2007) que sous l'hypothèse d'indépendance entre les interpolations des courbes et le plan de sondage, l'estimateur Horvitz-Thompson (5) est sans biais, et que l'estimation de sa variance est la somme des variances d'échantillonnage spatiales et des variances en chaque instant dues aux interpolations. En effet :

$$\hat{Var}(\hat{\tilde{C}}_{HT}(t)) = \sum_{i \in S} \sum_{j \in S} \left(\frac{\tilde{C}_i(t)}{\pi_i} \frac{\tilde{C}_j(t)}{\pi_j} (\pi_{ij} - \pi_i \pi_j) \right) + \sum_{i \in S} \frac{\sigma_i^2(t)}{\pi_i(t)} \tag{6}$$

Avec $\sigma_i(t)$ la variance du résidu $\epsilon_i(t)$ en chaque instant, que nous supposerons connue.

Le premier élément de cette somme peut être particulièrement bien maîtrisé, et la deuxième partie de ce calcul pourra être d'autant plus faible que nous pourrons, sur les courbes échantillonnés, utiliser un nombre de points de mesures plus important que dans le cas d'un recensement. Mais une difficulté supplémentaire apparaît à ce niveau. En effet, l'optimisation de l'échantillonnage temporel doit se fait sur l'échantillon spatial choisi. Aussi, l'erreur d'interpolation en un instant t d'une courbe de l'échantillon dépend non seulement de la technique d'interpolation utilisée, mais aussi de l'optimisation temporelle, et donc de l'échantillon sur lequel

l'optimisation est utilisée. Ainsi, nous voyons apparaître dans l'estimateur Horvitz-Thompson deux niveaux d'aléas :

1. Le plan de sondage
2. Le caractère aléatoire de l'interpolation des courbes conditionnellement à l'échantillon (et conditionnellement à la technique d'interpolation utilisée).

Des travaux théoriques sont en cours. Des approches combinant à la fois bootstrap sur plan de sondages et mouvements browniens pour aborder les erreurs d'interpolation sous contraintes (Abdessalem et al., 2005) seront abordés dans ce cadre.

4 Premières expérimentations

4.1 Méthodes d'échantillonnage

Afin de résumer les données provenant des capteurs, plusieurs solutions sont envisageables.

– Un échantillonnage purement temporel : tous les capteurs sont observés et la courbe prélevée de chaque capteur n'est pas complète. On sélectionne un ensemble minimal de données échantillonnées tout en s'assurant de la représentativité de ces données pour s'approcher de la courbe initiale. Pour celà, nous résolvons le problème suivant afin d'affecter de façon optimisée des pas d'échantillonnage à la totalité des courbes :

$$\text{Minimiser } \sum_{i=1}^{N} \sum_{j=1}^{m} (W_{ij} \times X_{ij})$$

sous les contraintes :

$$\begin{cases} X_{ij} = 0 \text{ ou } 1 & \\ \sum_{j=1}^{m} X_{ij} = 1 & \text{i de 1 à N} \\ \sum_{i=1}^{N} \sum_{j=1}^{m} (\left\lfloor \frac{p}{j} \right\rfloor \times X_{ij}) \leq s & \text{i de 1 à N} \end{cases}$$

Il s'agit d'un problème d'affectation de pas d'échantillonnage (granularité temporelle) aux différentes courbes en respectant les contraintes ci-dessus. Une courbe échantillonnée à un pas j signifie que nous effectuons un 'saut' de j entre deux points sélectionnés. Par exemple, $j = 2$ signifie que nous sélectionnons un point sur deux de la courbe. Les données prélevées sont ainsi équidistantes temporellement.
Une variable X_{ij} à 1 signifie que nous affectons le pas d'échantillonnage j à la courbe d'indice i . La deuxième contrainte du programme d'optimisation $\sum_{j=1}^{m} X_{ij} = 1$ impose une seule valeur de j par courbe (un seul pas d'échantillonnage). Enfin, la troisième contrainte signifie que le nombre de données à communiquer au système central ne doit pas dépasser le seuil imposé s. Chaque courbe doit être échantillonnée à un pas inférieur ou égal à m. Nous avons donc la possibilité de choisir un pas entier allant de 1 (nous récupérons tous les points de la courbe) à m (les points échantillonnés sont distancés par m). Nous calculons une matrice W_{n*m} de n lignes et m colonnes. Un élément w_{ij} de la matrice correspond à la somme des erreurs quadratiques obtenue si on applique un pas d'échantillonnage *j* à la courbe d'indice *i*. Pour résoudre ce problème, nous utilisons la méthode du simplexe appliquée aux problèmes linéaires à variables réelles. Le simplexe est couplé avec la méthode Branch And Bound (Séparation-Evaluation en français)

afin d'atteindre des variables entières. Le programme LP_Solve permet de résoudre des problèmes d'optimisation linéaires à variables réelles et/ou entières . Le lecteur peut se référer à (Gondran et al., 1979; lpsolve) pour des informations sur ces méthodes.

- Un échantillonnage purement spatial : il s'agit d'effectuer un plan de sondage simple sans remise afin de tirer un échantillon aléatoire et uniforme de capteurs. Le seuil imposé s détermine la taille de l'échantillon, celle-ci se trouve donc limitée par la bande passante disponible, et sa valeur est de : $n = \frac{s}{p}$. Chaque capteur i a la même probabilité d'inclusion $\pi_i = \frac{n}{N}$. Et chaque échantillon S qui peut être formé a la même probabilité de sortie $p(S) = \binom{N}{n}^{-1}$. Les courbes des capteurs ne faisant pas partie de l'échantillon sont estimées par la courbe moyenne de l'échantillon que nous 'calons' par rapport à la consommation globale de chaque capteur qui est une donnée connue.

- Echantillonnage spatio-temporel : Une troisième solution pour résumer les informations consiste à tirer aléatoirement un échantillon de taille n (échantillonnage spatial) et d'effectuer un échantillonnage temporel tel qu'il a été expliqué précédemment, soit au niveau de l'échantillon en cas de dépassement du seuil ($n * p \geq s$), soit au niveau des capteurs restant dans le cas contraire afin d'améliorer les estimations. Voici comment nous avons construit les courbes de consommation :
  - $n * p \geq s$: Les courbes de l'échantillon sont reconstruites par interpolation linéaire. les courbes ne faisant pas partie de l'échantillon sont estimées comme dans le cas purement spatial, par la courbe moyenne de l'échantillon que nous 'calons' par rapport à la consommation globale de chaque capteur.
  - $n * p < s$: Les courbes faisant partie de l'échantillon n'ont pas besoin d'être interpolées puisque nous récupérons la totalité des mesures. Pour chaque courbe ne faisant pas partie de l'échantillon, nous estimons les données entre deux instants échantillonnés temporellement par la courbe moyenne de l'échantillon prise entre ces deux instants.

4.2 Résultats

Nous avons expérimenté les approches décrites dans le paragraphe précédent grâce à un jeu de données composé de 1000 courbes de consommations électriques relevées à intervalle de 10 minutes pendant une journée (144 mesures par courbe). Nous avons expérimenté plusieurs valeurs de seuil s (nombre total de données que le serveur central peut accepter pendant une période donnée). Pour calculer les erreurs commises dans le cas de l'échantillonnage spatial de capteurs, nous avons effectué des simulations de Monte-Carlo (100 simulations).

Les résultats sont donnés sous forme de courbes dans la figure 1. Nous avons calculé la moyenne (ST), le maximum (STmax) et le minimum (STmin) des erreurs quadratiques commises lors des simulations dans le cas de l'échantillonnage Spatio-Temporel, et nous avons calculé les erreurs quadratiques globales (T) dans le cas de l'échantillonnage purement Temporel. L'abscisse dans les figures correspond au taux de compression appliqué à l'ensemble des points des courbes de consommation : un taux de compression de 2 signifie que nous récupérons $s = \frac{(144*1000)}{2}$ points. Nous avons fait varier la taille de l'échantillon n dans le cas de l'échantillonnage spatio-temporel.

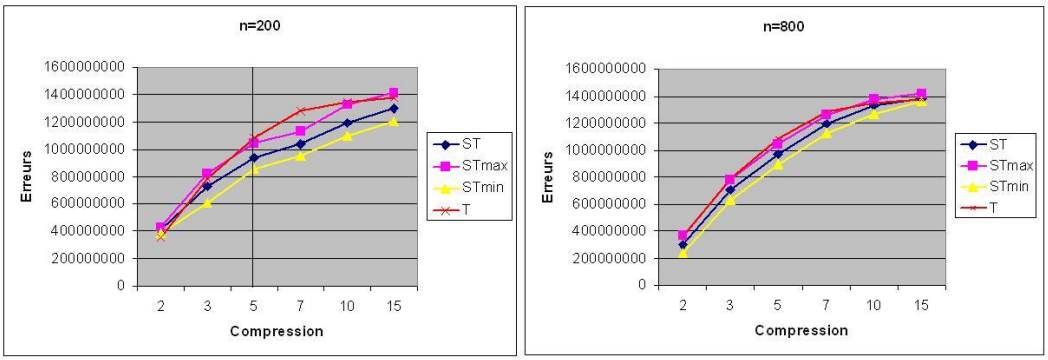

FIG. 1 – *Echantillonnage spatio-temporel Vs. Echantillonnage purement temporel*

Un échantillonnage purement spatial est donc appliqué aux courbes à gauche de la ligne verticale sur la figure correspondant à une taille de l'échantillon $n = 200$ (mode *non saturé*). A droite de la ligne, la taille n implique un dépassement du seuil, nous avons donc appliqué un échantillonnage spatio-temporel, ainsi que pour la figure de droite correspondant à une taille de l'échantillon de $n = 800$ (mode *saturé*). Nous remarquons que l'échantillonnage spatio-temporel permet de réduire en moyenne les erreurs commises par rapport à un échantillonnage purement temporel. Ce dernier s'avère meilleur dans le cas d'un échantillon de petite taille et un faible taux de compression. Toutefois, les deux courbes STmax correspondant au maximum des erreurs et T correspondant aux erreurs par échantillonnage purement temporel, se confondent et ceci pour toute valeur de n et tout taux de compression. A noter que la méthode de tirage effectuée est un échantillonnage aléatoire uniforme, nous envisageons d'expérimenter d'autres méthodes telle que l'échantillonnage stratifié, avec comme variables de stratification des données commerciales (équipements des clients, tarification,...), données géographiques, ou encore le niveau de consommation électrique.

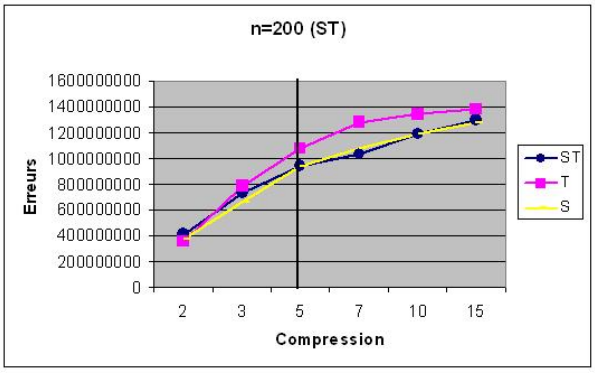

FIG. 2 – *Echantillonnage purement spatial Vs. Echantillonnage purement temporel Vs. Echantillonnage spatio-temporel de taille n=200*

La figure 2 montre les erreurs correspondant à l'échantillonnage purement spatial (S), l'échantillonnage purement temporel (T) et l'échantillonnage spatio-temporel (ST) avec une taille d'échantillon de $n = 200$. Nous avons expérimenté plusieurs valeurs de taux de compression. La taille de l'échantillon spatial dépend de ce dernier. Un taux de compression de 2 signifie que nous tirons un échantillon aléatoire uniforme de capteurs de taille $n = \frac{1000}{2}$. Les courbes (S) et (ST) correspondent aux courbes moyennes des erreurs quadratiques commises lors des simulations. La figure montre clairement que l'échantillonnage spatial permet d'estimer mieux la courbe de consommation globale que l'échantillonnage purement temporel et celà pour presque tous les taux de compression. Nous remarquons aussi que l'échantillonnage purement spatial a un comportement plus ou moins similaire à l'échantillonnage spatio-temporel avec une légère amélioration pour les faibles taux de compression. Les erreurs d'interpolation sont donc plus importantes que les erreurs d'échantillonnage dans le cas de nos courbes d'expérimentation.

5 Conclusion

Nous avons proposé une approche d'échantillonnage spatial de capteurs distribués couplée à un échantillonnage temporel. Nos premières expérimentations ont montré l'intérêt de cette approche par rapport à un échantillonnage temporel exhaustif (appliqué à tous les capteurs). Nous avons aussi montré que l'échantillonnage spatial s'avère meilleur pour estimer une courbe de consommation globale. Nous envisageons d'expérimenter ces trois méthodes pour estimer des courbes individuellement ou sur des petits domaines et comparer leur comportement dans ce contexte. Toutefois, des problèmes se posent quant à l'estimation de la précision avec ou sans dépendance entre le plan de sondage et les lissages utilisés. Nous travaillons afin de démontrer de façon théorique le mérite d'un échantillonnage spatio-temporel par rapport à un échantillonnage simplement spatial ou simplement temporel.

Une perspective de ce travail est d'appliquer une approche panel aux flux de données provenant de capteurs. On définit un panel comme un échantillon où les individus sont interrogés au moins deux fois (Ardilly, 2006). A la différence de l'échantillonnage qui fournit une information ponctuelle, les panels ont pour vocation de mesurer l'évolution de cette même information au cours du temps. En général, une partie du panel est renouvelée selon un intervalle de temps fixé, pour prendre en compte l'évolution de la population ou pour faire face aux non réponses ou à la lassitude des individus. Cependant, dans le cas des données provenant de capteurs, nous n'avons pas de phénomène de lassitude. De plus, le renouvellement du panel n'engendre pas de coût supplémentaire, puisque les capteurs sont déjà installés chez les clients. Par conséquent, nous ferons évoluer dynamiquement le panel selon les variations des données observées, de l'évolution des besoins d'analyse et des évolutions des données contractuelles (nouveaux clients, changement d'équipements, déménagements...).

Références

Abdessalem, T., Decreusefond, L., et Moreira, J. (2005). Probabilistic measurement of uncertainty in moving objects databases. In BDA'2005, 21èmes Journées de Bases de Données Avancées.

Aggarwal Charu C. (2007). Data streams : models and algorithms, IBM Watson Research Center, Springer Editions.

Ardilly P. (2006). Les techniques de sondage, Editions Technip.

Babcock B., Babu S., Datar M., Motwani R., Widom J., (2002). Models and issues in data stream systems. In Symposium on Principles of Database Systems, pp.1-16. ACM SIGACT-SIGMOD, 2002.

Chiky R. et Hébrail G. (2007). Echantillonnage optimisé de données temporelles distribuées pour l'alimentation des entrepôts de données. EDA (Entrepôt de Données et Analyse en ligne), Poitiers.

Dessertaine A. (2007). Echantillon et flux de données - estimations de courbes de consommation électrique à partir de données fonctionnelles. Colloque francophone sur les sondages 2007, Marseille.

Deville, J. C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. Annales de l'INSEE, p. 3-104.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. Survey Methodology, 25, 2, 193-203.

Golab L., Ozsu M.T., (2003). Data stream management issues - a survey, Technical report CS 2003-08, University of Waterloo, Waterloo, Canada, April 2003.

Gondran M., Minoux M., (1979). Graphes et algorithmes. Eyrolles, Paris 1979.

Jain A. et Chang E. Y. (2004). Adaptive sampling for sensor networks. In Proceeedings of the 1st international Workshop on Data Management For Sensor Networks : in Conjunction with VLDB 2004 Toronto, Canada.

http ://lpsolve.sourceforge.net/

Muthukrishnan S., (2005). Data streams : algorithms and applications, In Foundations and Trends in Theoretical Computer Science, Volume 1, Issue 2, August 2005.

Ramsay J.O. et Silverman B.W. (2005). Functional Data Analysis, New-York : Springer-Verlag

Tillé, Y. (2006). Sampling Algorithms, New-York : Springer-Verlag

Stonebraker M., Cetintemel U., Zdonik S. (2005). The 8 requirements of real-time stream processing, SIGMOD Records.

http ://www-db.stanford.edu/stream/

http ://www.streambase.com

Willett R., Martin A., et Nowak R. (2004). Backcasting : adaptive sampling for sensor networks. In Proceedings of the Third international Symposium on information Processing in Sensor Networks, Berkeley, California.

Summary

A growing number of real world applications deal with multiple streams of data produced by sensors: meteorolgy forecasting, military applications, electric consumption analysis,etc. Stream mining and analysis techniques employed in these applications have to be efficient in terms of space usage and process streams 'on the fly' as they are produced. In particular, if

the system is not able to process all data streams, summaries are built. This paper proposes a new approach for sampling a set of distributed data streams. Our approach is based on survey theory, functional data analysis and data stream algorithms. This approach will be studied in relation to a real application on electric power consumption measurement.