
	<p>RTE direction R&D Innovation & Ecole Centrale de Lille option DAD (Décision & Analyse de Données)</p> <p>Proposition de projet IMPACT 2015-2016</p> <p>« Calculs dynamiques de quantiles sur flux de données – algorithme P2 vs »</p>	
---	---	---

Sujet : *Tester les performances d'un algorithme de calcul itératif, dynamique de quantiles sans avoir à stocker et charger l'ensemble des observations et pouvant servir à la mise à jour de quantile au fil de l'eau dans un processus d'arrivée continue de nouvelles données. On se fondera sur un algorithme spécifique dit « The P² Algorithme » décrit dans l'article fourni en annexe de ce sujet. Il s'agit de tester cette méthode et d'en évaluer les performances, ses avantages et inconvénients, dans le logiciel d'analyse R et ceci sur un jeu de données réelles (transits et/ou tensions et/ou consommation et/ou production) fourni par RTE. Si le temps le permet, l'étudiant pourra confronter de manière critique cet approche à celles accessibles par des régressions de quantiles comme proposées dans le package R « quantreg ».*

Contexte et enjeux :

En tant que statistiques de répartition et de description de données, les quantiles sont des outils considérés comme standard et largement maîtrisés. La volumétrie s'accroissant dans le cadre des analyses dites « Bigdata » et les contraintes opérationnelles consistant à travailler sur des flux de données arrivant continuellement ou presque soulève des défis sur des tâches aussi simples que... calculer des quantiles. Comment les calculer efficacement, rapidement, à moindre coût et en minimisant le besoin de stocker et d'accéder en permanence à l'intégralité des observations si tant est qu'elles aient été historisées par ailleurs.

RTE est confronté à ces sujets dans le cadre de l'analyse des données du système électrique français (consommations, productions, données réseaux). On retrouvera le besoin de calculer des quantiles dans le cadre de l'analyse de risque, de monitoring de variables systèmes, de simulations de situations réalistes, etc. L'enjeu est la maîtrise de ces phases de calculs au « meilleur » coût d'autant plus lorsque l'on est contraint en temps d'exécution (dans le cadre de processus opérationnels en exploitation dans les métiers de RTE) et que la taille des données est critique (jeux de données pluriannuels, sur plusieurs milliers à dizaines de milliers de variables, au pas minute par exemple).

Pour ce projet, on se fondera sur la publication en annexe de ce sujet : "The P² Algorithm for Dynamic Calculation of Quantiles and Histograms Without Storing Observations" de Raj JAIN et Imrich CHLAMTAC, "Communications of the ACM - October 1985 - volume 28 - number 10".

On utilisera préférentiellement le logiciel opensource d'analyse de donnée R et RTE fournira comme cas d'application un jeu de données de consommations électriques élémentaires et/ou de productions sur le système français et/ou de transits sur ses liaisons. Des essais sur données simulées pourront être menés préalablement éventuellement. Si le temps le permet, l'étudiant pourra confronter de manière critique cet approche à celles accessibles par des régressions de quantiles comme proposées dans le package R « quantreg ».

Encadrant RTE :

Samir ISSAD
Ingénieur Responsable d'études / Engineer
RTE - R&D Innovation
Immeuble LE COLBERT
9, Rue de la Porte de Buc , BP 561
78005 VERSAILLES CEDEX
samir.issad@rte-france.com
Tel : +33 (0)1.39.24.40.16
Fax : +33 (0)1.39.24.41.75