Sélection optimale de capteurs de référence pour le stockage de données spatialement corrélées

Antoine CRINIÈRE[1], Aline ROUMY[2], Thomas MAUGEY[2], Michel KIEFFER[3], Jean DUMOULIN[1,4]

[1]Equipe I4S Inria Rennes
Campus de Beaulieu, 263 Avenue Général Leclerc, 35042 Rennes

[2]Equipe SIROCCO Inria Rennes
Campus de Beaulieu, 263 Avenue Général Leclerc, 35042 Rennes

[3]L2S, UMR CNRS 8506; CentraleSupelec; Univ. Paris-Sud

[4]IFSTTAR, COSYS-SII, Route de Bouaye - CS4, F-44344, Bouguenais, France
antoine.criniere@inria.fr

Résumé – Les nouvelles politiques urbaines s'intéressent de près aux villes intelligentes hautement instrumentées faisant naître ainsi des problématiques de gestion et stockage d'un large volume de données émanant d'un nombre croissant de sources. Cette étude détaille une méthode de compression par codage prédictif de données multisources spatialement corrélées, basée sur l'expression de sources de références et l'utilisation du Krigeage comme méthode de prédiction

Abstract – Highly instrumented Smart-cities, which are now a common urban policies, are facing problems of management and storage of a large volume of data coming from an increasing number of sources. This study presents a data compression method by predictive coding of spatially correlated multi-source data based on reference selection and prediction by Kriging

1 Introduction

Avec l'avancée des technologies de l'information et de la communication, émergent les concepts de capteurs intelligents et de territoires connectés [1]. Ces dernières années ont vu le développement de réseaux de capteurs à l'échelle mondiale et l'avènement de services connectés permettant l'agrégation et la diffusion des données issues de ces réseaux [2]. De plus les nouvelles politiques urbaines s'intéressent de près aux villes intelligentes hautement instrumentées, par exemple le projet SenseCity [3]. Ces nouveaux développements font apparaître la multiplicité des capteurs et font naître des problématiques de gestion et stockage d'un large volume de données émanant d'un nombre croissant de sources, à l'image des données environnementales [4]. Dans ce contexte le projet Intercom [1] s'intéresse à de nouvelles méthodes de compression de données corrélées et notamment aux problématiques soulevées par les territoires connectés. Dans cette optique, l'étude présente le développement d'une méthode de compression par codage prédictif de données multi-sources spatialement corrélées, basée sur l'expression de sources de références et l'utilisation du Krigeage comme méthode de prédiction. Un algorithme de sélection optimal des références est introduit, la méthode est alors appliquée à un ensemble de capteurs météo répartis sur le territoire métropolitain.

1. http://www.intercom.cominlabs.ueb.eu Ce projet est financé par le laboratoire d'excellence CominLabs ANR-10-LABX-07-01.

2 Compression de données

2.1 Modèle de données

Chaque capteur est modélisé par une source aléatoire X qui génère une suite de variables aléatoires indépendantes et identiquement distribuées i.i.d. de loi $p(x) = \mathcal{P}(X = x)$ et d'alphabet $\mathcal{X}$. Les mesures acquises par ce capteur sont, avec ce modèle, les réalisations de ces variables aléatoires. Le taux optimal de compression sans perte de cette source est donné par l'entropie [5], équation 1, qui représente le nombre de bits moyens nécessaires pour représenter un échantillon d'une source X

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) \tag{1}$$

et est exprimé en nombre de bits par échantillon. Dans notre étude, nous considérons un ensemble de L capteurs prenant leur indice dans un ensemble $\mathcal{L}$. Lorsque les mesures de tous les capteurs sont encodées conjointement, le débit minimal de compression $\mathcal{R}_{\text{joint}}$ est donné par l'entropie jointe :

$$\mathcal{R}_{\text{joint}} = H(X_1, ..., X_L) = -\sum_{(x_1,...,x_L)\in\mathcal{X}^L} p(x_1, ..., x_L)\log_2 p(x_1, ..., x_L) \tag{2}$$

Ce débit est exprimé en bits moyen par acquisition de l'ensemble des sources, on parlera ici de *bits par capture*.

Les capteurs ayant des mesures spatialement corrélées, il est important de compresser les sources conjointement.

$$\mathcal{R}_{\text{séparé}} = \sum_{l=1}^{L} H(X_l) \geq \mathcal{R}_{\text{joint}} = H(X_1, X_2, ..., X_L) \quad (3)$$

2.2 Schéma de compression

Parmi les différentes méthodes de compression de données climatiques [6][7], et notamment de données radar [8], nombres d'entre elles exploitent la corrélation temporelle des données *d'un capteur* [9]. Ici, nous proposons une méthode de codage sans perte et multisources, figure 1, qui permet au contraire d'exploiter la corrélation *entre les capteurs*. Ce schéma est *prédictif* afin d'exploiter la corrélation entre les sources tout en maîtrisant la complexité de l'encodeur. En effet, cette méthode ne nécessite ni d'estimer ni de stocker la loi jointe des sources.

Cette méthode consiste à sélectionner un ensemble $\mathcal{K} \subset \mathcal{L}$ de K sources de référence puis la prédiction des sources de l'ensemble $\mathcal{L} \setminus \mathcal{K}$ à partir des références. Cette étape de prédiction est corrigée par l'envoi de résidus $\Psi_{\mathcal{L}\setminus\mathcal{K}}$. Ainsi le débit de compression atteint par notre schéma est :

$$\mathcal{R} = \sum_{i \in \mathcal{L}\setminus\mathcal{K}} H(\Psi_i) + \sum_{i \in \mathcal{K}} H(X_i) \quad (4)$$

Pour l'étape de prédiction, différentes méthodes existent dans la littérature pour interpoler des données à partir de données de référence. L'ensemble de nos capteurs étant distribué spatialement une méthode d'interpolation de données régionalisées apparaît justifiée. Le Krigeage est une méthode géostatistique d'interpolation spatiale qui permet l'estimation linéaire non-biaisée optimale (car garantissant le minimum de variance) d'une variable régionalisée Y [10]. Dans la littérature, le krigeage a déjà été utilisé pour la compression de données [11] qui sont échantillonnées de manière régulière et dense, comme une image. Ici, nous nous considérons des données échantillonnées de manière irrégulière. Néanmoins, la méthode de krigeage est similaire : elle permet le calcul les poids λ_i de la combinaison linéaire suivante :

$$\hat{y}(s_0) = a + \sum_{i=1}^{I} \lambda_i y(s_i) \quad (5)$$

où I est le nombre de capteurs, $y(s_i)$ une mesure à la position s_i et $\hat{y}(s_0)$ l'estimation de la variable régionalisée en la position s_0. L'intérêt du Krigeage en géostatistique réside dans le fait qu'il tient compte de la dépendance spatiale de Y par l'intermédiaire du variogramme [12]. Le variogramme, équation 6, représente le degré de dépendance spatiale d'un champ aléatoire et ne dépend que de l'interdistance h entre deux capteurs et non de la position de ceux ci. Lors de l'analyse le variogramme expérimental est calculé à partir des données, équation 7 où $D(h)$ est le nombre de paires de points satisfaisant $h = (s_i - s_j)$. Dans le cadre de l'analyse krigeante le variogramme doit être continu c'est pourquoi il est nécessaire d'utiliser un modèle dont les paramètres sont estimés à partir des données, équation 8.

$$\gamma(h) = \frac{1}{2}\left[Var(Y(s_i + h) - Y(s_i))\right] \quad (6)$$

$$\gamma_{expe}(h) = \frac{1}{2D(h)} \sum_{D(h)} \left(y(s_i) - y(s_j)\right)^2 \quad (7)$$

$$\gamma(h) = c\left\{1 - \exp\left(\frac{h^\alpha}{r^\alpha}\right)\right\} \quad (8)$$

où $\alpha = 1.5$ et $[c; r]$ respectivement le seuil et la portée du variogramme. Le modèle utilisé est dit *stable*, son choix a été motivé par son large champ d'application ainsi que par la spécificité de cette étude qui ne cherche pas à reconstruire fidèlement un champ de donnée mais à générer des résidus. Avec l'expression analytique du variogramme, les poids de l'estimateur sont calculés en résolvant $\Lambda = \Gamma^{-1}(\gamma_0 + \ell)$, avec $\Gamma_{i,j} = \gamma(||s_i - s_j||_2)$, γ_0 le vecteur où la composante i est $\gamma(||s_i - s_0||_2)$ et ℓ un multiplicateur de Lagrange, ce qui permet de résoudre l'équation du Krigeage ordinaire

$$\hat{y}(s_0) = \Lambda^t Y \quad (9)$$

où $Y^t = [y(s_1), ..., y(s_I)]$. Le krigeage est traditionnellement utilisé pour estimer des mesures qui ne sont pas connues. Dans cette étude au contraire, le krigeage est utilisé pour estimer, à partir des mesures de référence ($\mathcal{K}$), les mesures certes connues des autres capteurs, mais non sélectionnées comme référence ($\mathcal{L} \setminus \mathcal{K}$). Une fois cette estimation faite, seul le résidu est transmis. Ainsi pour chaque source $X_i \in \mathcal{L} \setminus \mathcal{K}$ à l'instant $n \in [0, 1, ..., N]$ les résidus sont exprimés équation 10, où $\mathbf{X}_\mathcal{K} = \left[x_n^1, ..., x_n^K\right]$.

$$\begin{aligned} \Psi_{i,n} &= \hat{x}_{i,n} - x_{i,n} \quad \forall i \in \mathcal{L} \setminus \mathcal{K} \\ &= \Lambda_i^t \mathbf{X}_\mathcal{K} - x_{i,n} \quad \forall i \in \mathcal{L} \setminus \mathcal{K} \end{aligned} \quad (10)$$

2.3 Sélection optimale des capteurs de référence

Dans cet article, nous proposons d'optimiser le schéma de compression et cherchons l'ensemble $\mathcal{K} \in \mathcal{L}$ qui minimise le débit de compression $\mathcal{R}$. Plus précisément, cet ensemble optimal est solution de 11.

$$\mathcal{K}^* = \arg\min_{\mathcal{K}} \sum_{i \in \mathcal{L}\setminus\mathcal{K}} H(\Psi_i) + \sum_{i \in \mathcal{K}} H(X_i) \quad (11)$$

Le placement optimal des capteurs a déjà été traité dans un cadre, où le capteur peut être déplacé (s'il est mobile) [15] ou placé (s'il est fixe) [14] sur une grille continue. Ici, le choix des positions des capteurs est faite parmi un ensemble fini et prédéterminé, ce qui nécessite un nouvel algorithme d'optimisation. Une recherche exhaustive pouvant être trop coûteuse (le nombre de solutions à tester est $\binom{N}{K}$ et chaque calcul de débit nécessite un krigeage et donc une inversion de matrice), nous

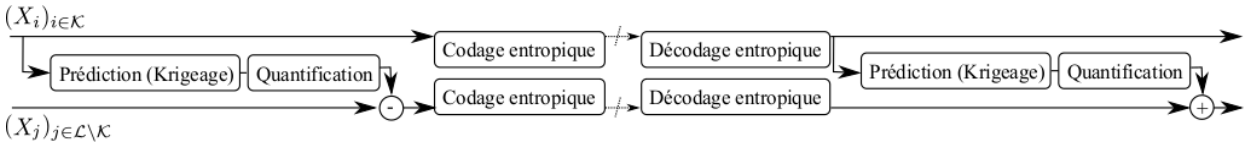


FIGURE 1 – Schéma de compression sans perte utilisé

Algorithme 1 Estimation séquentielle de $\mathcal{K}^*$

```
for i = 1 ... Lmax do
    for j ∈ L do
        if CapteurNonTraité then
            K_j = [K_{i-1}; j]
            Rtemp_j = CODAGE(N, K_j, L, X)        ▷ Figure 1
        end if
    end for
    K_i = arg min_{K_j} (Rtemp)                   ▷ Ajout du minimum local
    Rmin_i = min Rtemp
end for
K* = arg min_{K_i} (Rmin)
```

proposons un algorithme séquentiel sous optimal, algorithme 1.

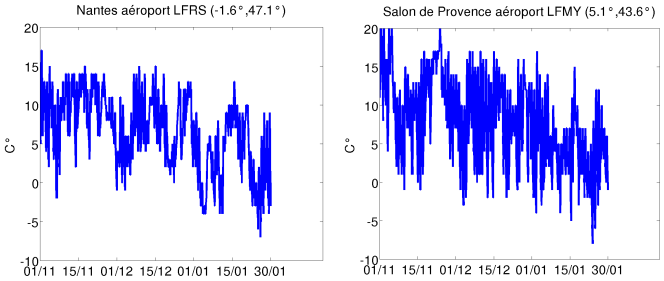
Cet algorithme identifie un jeu de capteurs $\mathcal{K}_i$ qui minimise localement le débit et l'utilise comme initialisation de la boucle $i+1$, $Lmax$ est la taille maximale du jeu de capteur. Soit un nombre de calculs de débit égal à :

$$\sum_{i}^{Lmax} L-(i-1) \approx \frac{L^2}{2} \quad ; \quad \textbf{if } Lmax = L \tag{12}$$

3 Résultats

3.1 Données utilisées

Dans le cadre de cette étude, trois mois de données météorologiques acquises à $5e^{-4}$Hz et reparties sur l'ensemble du territoire métropolitain ont été récupérés au format METAR par l'intermédiaire du réseau MESONET [2]. Nous nous intéressons aux mesures de températures, figure 2.



(a) (b)

FIGURE 2 – Températures mesurées de Novembre 2016 à Janvier 2017 à Nantes (a) et Salon de Provence (b)

A partir de ces données il est possible d'estimer pour chaque source une distribution $\hat{p}(x)$. Cette distribution permet ensuite de proposer un calcul empirique de l'entropie à partir de N échantillons $\{x_n\}_{1\leq n\leq N}$:

$$\hat{H}(X) = -\frac{1}{N}\sum_{1\leq n\leq N} \log_2(\hat{p}(x_n)) \tag{13}$$

La figure 3 présente une carte des entropies empiriques calculées pour chacune des sources dont nous disposons, L=70, présentant $N = 4200$ réalisations.

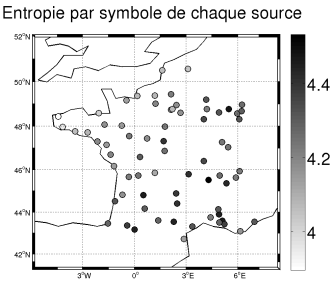


FIGURE 3 – Entropie de chacunes des sources utilisés

Il est intéressant de noter ici que les sources situées dans des zones climatiques particulières, comme des régions montagneuses ou bénéficiant d'un ensoleillement plus conséquent (S, S/E, E), font apparaître une plus grande variabilité de la mesure, figure 2(b), et donc une entropie plus importante. Dans un souci d'amélioration du temps calcul et après analyse des signaux présentés figure 2, le variogramme est calculé tous les 7 jours. Cette spécificité implique une diminution des gains en débit de l'algorithme, à l'avenir le variogramme devra être calculé à chaque instant. Chaque calcul su variogramme implique la transmission de deux flottants au décodeur.

3.2 Analyse des résultats

La figure 4 présente l'évolution du débit $\mathcal{R}$ pour un nombre maximal de capteurs utilisés comme référence de $Lmax = \frac{L}{2}$. Le débit y est comparé au débit initial pour lequel $\mathcal{K} = \mathcal{L}$.

Premièrement il apparaît que le débit chute rapidement vers l'optimum qu'il atteint pour un nombre relativement restreint de sources de référence ($\approx 15\%$). Ce qui apparaît encourageant pour de futures études sur un nombre conséquent de sources. Le taux de compression de cette méthode est de $\tau = \frac{R^*}{Base} = 0.7$ soit un gain en volume de 0.3.

L'analyse de la figure 5 fait apparaître à l'optimum un taux de 60% de résidus transmis. La concavité de la courbe de débit des références montre que les premières sources de références

2. https://mesonet.agron.iastate.edu

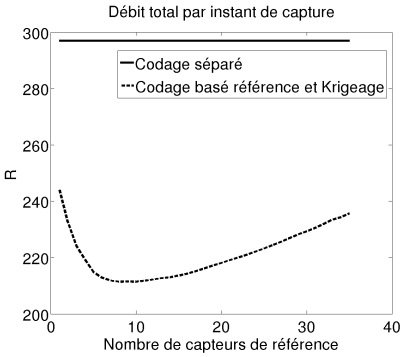


FIGURE 4 – Courbes de débit en codage séparé

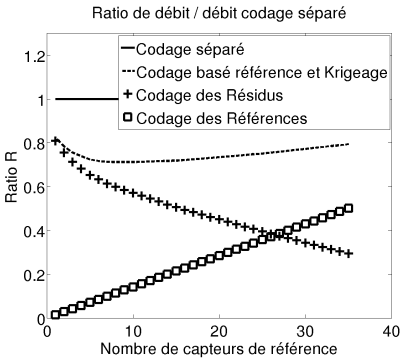


FIGURE 5 – Ratio de débit / débit codage séparé

sélectionnées sont celles qui ont une grande entropie.

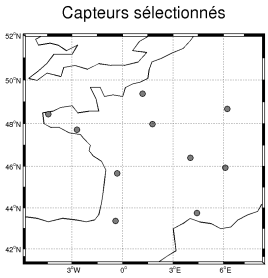


FIGURE 6 – Ensemble des capteurs de référence sélectionnés par l'algorithme 1

La figure 6 présente l'emplacement des capteurs de références sélectionnés. A première vue ces capteurs ne semblent pas être dépendants de la granularité spatiale des mesures et apparaissent répartis équitablement.

Enfin la figure 7 présente l'entropie des résidus et l'information mutuelle entre ceux-ci et la prédiction, rapportée à l'entropie de la mesure. L'entropie des résidus apparaît relativement faible par rapport à l'entropie des mesures, figure 3. Elle est néanmoins, à l'image de cette dernière, plus importante dans les zones présentant une forte variabilité climatique. Le ratio d'information mutuelle montre une faible dépendance entre résidus et prédictions confortant ainsi la pertinence de la méthode de prédiction proposée.

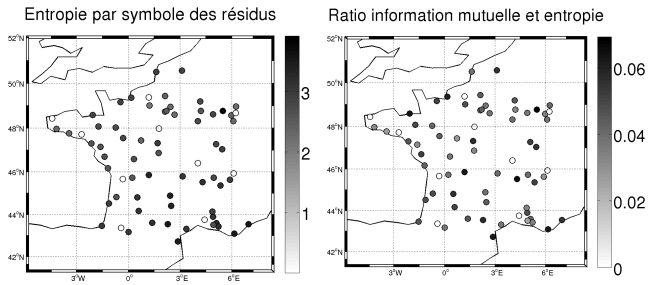


FIGURE 7 – Entropie des résidus à gauche et information mutuelle des résidus avec la prédiction rapportée à l'entropie de la mesure à droite

4 Conclusions

Cette étude a présenté une méthode de sélection optimale de capteurs référence pour l'utilisation d'un codeur entropique basé sur une prédiction par Krigeage. Après avoir introduit la méthode d'interpolation spatiale ainsi que le calcul de débit de stockage un algorithme d'optimisation est proposé. Cette méthode identifie un jeu de sources $\mathcal{K}$ optimal au sens du débit de stockage. Cet optimum ne nécessite qu'un faible nombre de sources de références et permet un gain de stockage d'au moins un tiers. Les résultats des calculs d'information mutuelle entre résidus et prédictions sont néanmoins légèrement sensibles à la variabilité climatique de certaines des zones considérées.

Références

[1] K.A. Delin. *The Sensor Web : A Macro-Instrument for Coordinated Sensing.* Sensors, Vol. 2, 2012.

[2] C. Reed *et al. Sensor web enablement : Overview and high level architecture. White Paper 07-165r1* Open Geospatial Consortium, 2013.

[3] F. Derkx. *et al. The Sense-City project.* Vibrations, Shocks and Noise XVIII symposium, 2012.

[4] S. Witt *et al. Managing large data volumes from scientific facilities.* ERCIM News Special theme : Big Data, 2012.

[5] C. Shannon. *A mathematical theory of communication. Bell System Technical* Journal, 27, 1948

[6] J. Gertz. *The Weather-Huffman Method of Data Compression of Weather Images* Lincoln Laboratory, 1997

[7] C. Steffen *et al. Weather Data Compression.* 19th Conference on IIPS, 2003.

[8] W. Ai *et al. Efficient Lossless Compression of Weather Radar Data.* International Journal ECEGGE Vol :3, No :8, 2009

[9] D. Pan *et al. Lossless differential compression of weather radar data in universal format using motion estimation and compensation.* 26th Conference on IIPS, 2010.

[10] G. Matheron *Les variables régionalisées et leur estimation.* Masson, Paris, 1965.

[11] A. Yfantis *et al. Image compression and kriging.* Quantitative Geology and Geostatistics, 6, 1996.

[12] N. Cressie. *Statistics for Spatial Data.* A Wiley-Interscience Publication, 1993.

[13] R. Webster et M. Oliver. *Geostatistics for Environmental Scientists (2nd ed.).* John Wiley & Sons, Ltd, 2007.

[14] A. Min et K. Shin *Joint optimal sensor selection and scheduling in dynamic spectrum access networks.* Ieee Transactions On Mobile Computing, 2013.

[15] A. Kahn, J. Marzat, H. Piet-Lahanier et M. Kierffer *Global extremum seeking by Kriging with a multi-agent system.* IFAC, 2015.