
Davide Andrea Guastella, Valérie Camps, Marie-Pierre Gleizes

Estimation d'informations environnementales avec le système HybridIoT : un cas d'étude sur la ville de Toulouse

Volume 5, nº 1 (2024), p. 63-91.

https://doi.org/10.5802/roia.65



*La Revue Ouverte d'Intelligence Artificielle est membre du Centre Mersenne pour l'édition scientifique ouverte*
*www.centre-mersenne.org*
*e-ISSN : 2967-9672*

**Revue Ouverte d'Intelligence Artificielle**
Volume 5, n° 1, 2024, 63-91

# Estimation d'informations environnementales avec le système HybridIoT : un cas d'étude sur la ville de Toulouse

**Davide Andrea Guastella[a], Valérie Camps[b], Marie-Pierre Gleizes[b]**

[a] Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgique
*E-mail :* davide.andrea.guastella@ulb.be

[b] Institut de Recherche en Informatique de Toulouse, Université Toulouse III - Paul Sabatier, 31062 Toulouse Cedex 9, France
*E-mail :* valerie.camps@irit.fr, marie-pierre.gleizes@irit.fr.

Résumé. — La ville intelligente s'intéresse à l'amélioration de la qualité de vie de ses habitants. De nombreux capteurs *ad hoc* nécessitent alors d'être déployés pour connaître l'état de l'environnement dans lequel les activités humaines se déroulent. Si ces capteurs sont souvent bon marché, leurs coûts d'installation et de maintenance augmentent rapidement avec leur nombre. La problématique adressée dans ce papier consiste à estimer des informations environnementales dans le cas où des capteurs physiques ne sont pas disponibles, pour limiter les coûts engendrés par l'installation et la maintenance de capteurs supplémentaires.

Le système HybridIoT permet d'estimer les valeurs environnementales manquantes dans des réseaux de capteurs à grande échelle à l'aide de mécanismes distincts : (*i*) une estimation endogène à partir d'un historique de données, (*ii*) une estimation endogène à l'aide de capteurs voisins homogènes et (*iii*) une estimation exogène. Ce papier s'intéresse à une amélioration de la technique d'estimation à l'aide de capteurs voisins homogènes (*ii*). Plus concrètement, notre contribution est triple : la définition d'une nouvelle approche géospatiale pour estimer des valeurs manquantes dans des environnements à grande échelle, l'évaluation de cette méthode géospatiale pour estimer des valeurs environnementales sur la ville de Toulouse, et enfin les premières avancées sur le déploiement du système dans le cadre du Groupement d'Interêt Scientifique (GIS) neOCampus.

Mots-clés. — Ville intelligente, systèmes multi-agents coopératifs, estimation de données manquantes.

## 1. Introduction

La ville intelligente (*smart city*) a pour objectif de répondre à des problèmes d'optimisation de ressources afin de permettre de meilleures interactions entre citoyens. Elle s'intéresse à plusieurs aspects de la société urbaine ; nous nous focalisons ici sur l'aspect technologique, à savoir l'utilisation de technologies récentes pour analyser les

données de l'environnement urbain afin d'améliorer les services offerts aux citoyens, mais aussi leur qualité de vie.

Pour atteindre ces objectifs d'amélioration, il est nécessaire de pouvoir observer de façon ponctuelle et continue l'environnement par le biais de capteurs, sachant que l'idéal serait de parvenir à un compromis entre la précision des informations observées et les coûts d'installation et de maintenance. Afin de réduire ces coûts, supportés par les collectivités, nous proposons de leur fournir des moyens technologiques permettant de combler le manque de capteurs, autrement dit de leur fournir des estimations de données environnementales dans des points non couverts par des capteurs sans coût supplémentaire. Cette proposition est l'objectif du système HybridIoT dont nous allons présenter les principes. Concrètement, HybridIoT permet de définir une infrastructure IoT *hybride*, dans laquelle des capteurs physiques ainsi que des capteurs virtuels fonctionnent conjointement et de manière transparente pour l'utilisateur du système, afin de fournir des informations en des points précis de l'environnement. L'utilisation d'agents logiciels agissant en tant que capteurs virtuels permet de fournir des informations à des endroits non couverts par des capteurs physiques, sans installation de nouveaux capteurs.

Le reste du présent document est organisé comme suit : la section 2 positionne ce travail par rapport aux principales techniques d'estimation de valeurs manquantes dans différents domaines d'application. La section 3 décrit le système HybridIoT, qui permet d'estimer des valeurs environnementales manquantes dans des zones non couvertes par des capteurs. La section 4 présente les résultats d'une nouvelle méthode géospatiale pour l'estimation de valeurs environnementales obtenus à partir d'une étude de cas portant sur la ville de Toulouse. Dans la section 5, nous résumons les activités en cours dans le cadre du déploiement du système HybridIoT dans un contexte opérationnel réel. Dans la section 6, nous concluons notre travail et indiquons quelques perspectives de travail.

## 2. Positionnement et contexte

La ville intelligente joue un rôle clé dans la transformation des contextes urbains en améliorant différents aspects de la vie de ses citoyens, tels que l'environnement, le transport, la santé, l'énergie et l'éducation. Le principal défi de la ville intelligente est de faire un usage intensif des données acquises à l'échelle urbaine. Cette quantité de données augmente considérablement aujourd'hui, principalement en raison de l'accessibilité et du faible coût des dispositifs capables d'acquérir des données environnementales. Un grand nombre de dispositifs de détection disséminés dans le contexte urbain génère un énorme volume de données, généralement appelé *big data*, qui est au cœur des services rendus par l'IoT. L'utilisation du *big data* offre à la ville la possibilité d'obtenir des informations précieuses à partir d'une quantité considérable de données collectées à partir de diverses sources [13]. Les données acquises à partir de dispositifs GPS peuvent être utilisées pour surveiller les conditions de circulation afin de réduire les retards, les bouchons et les accidents. Dans les bâtiments, les données

acquises à partir de capteurs (luminosité, température, etc.) et les habitudes des utilisateurs peuvent être utilisées pour contrôler les systèmes de chauffage, de ventilation et de climatisation afin d'optimiser la consommation d'énergie tout en garantissant une bonne qualité de vie aux utilisateurs.

Dans le but d'optimiser la gestion des ressources et d'améliorer les services des villes, il est nécessaire non seulement de collecter une grande quantité de données, mais aussi d'extraire des connaissances utiles pour atteindre les objectifs de « smartness ». Dans ce contexte, l'Intelligence Artificielle (IA) peut contribuer à l'analyse, l'extraction de connaissances et les raisonnements sur de gros volumes de données afin que les acteurs de la ville intelligente puissent décider des actions appropriées à mener. L'utilisation conjointe de l'IoT et l'IA permet de passer des systèmes d'objets connectés à des systèmes d'intelligence connectée.

L'estimation de valeurs manquantes dans des ensembles de données est apparue dans les années 1970 en tant que technique pour traiter l'incomplétude des données. Cette tâche est nécessaire dans le domaine des villes intelligentes pour plusieurs raisons : la présence de données incorrectes, des capteurs qui ne fonctionnent pas (ou qui ne sont pas disponibles) et/ou la nécessité d'estimer et de déduire des données avec plus de précision à l'aide des données disponibles [18]. Bien que le manque de données soit prévisible dans de nombreuses applications, l'estimation des données manquantes est un défi car, dans de nombreux domaines, elle doit être réalisée en temps réel et à la demande. Étant donné que les informations estimées peuvent être utilisées dans les processus décisionnels, les techniques d'estimation doivent produire des résultats précis. Par conséquent, ces estimations doivent être aussi proches que possible des valeurs réelles afin de pouvoir prendre des décisions efficaces.

Le système HybridIoT tente de lever trois verrous : la prise en compte de la dynamique imprévisible de l'environnement, la décentralisation du calcul et l'hétérogénéité des entités présentes dans l'environnement. Pour cela, la solution proposée adresse simultanément trois propriétés : l'ouverture, l'hétérogénéité et le passage à l'échelle. L'ouverture permet au système de continuer à fonctionner sans aucune reconfiguration, même lors de l'apparition ou la disparition de capteurs. L'hétérogénéité permet l'intégration d'informations ayant un type et/ou une échelle de mesure différents, provenant de plusieurs sources de données. Le passage à l'échelle indique la possibilité de déployer le système dans des contextes urbains tels que les villes ou les régions. Nous avons passé en revue les principales techniques de l'état de l'art traitant de l'estimation de valeurs manquantes dans différents domaines d'application [6] et nous les avons évaluées par rapport à leur capacité d'adresser les trois propriétés souhaitées (tableau 2.1).

Nous avons utilisé quatre indicateurs pour décrire les points forts et les points faibles de chaque méthode décrite : (++) une propriété a été discutée et les auteurs présentent une solution pour la traiter, (+) une propriété a été discutée et traitée mais les auteurs n'ont pas fourni une description détaillée de la solution, (–) la propriété a été mentionné mais pas traité, (– –) la propriété n'a été ni mentionnée ni traitée.

TABLE 2.1 – Comparaison des solutions de l'état de l'art pour l'estimation des informations manquantes.

| **Technologie** | **Auteurs** | **Domaine** | **Ouverture** | **Hétérogénéité** | **Passage à l'échelle** |
|---|---|---|---|---|---|
| Régression | Hasenfratz *et al*. (2015) [12] | Environnement | - - | ++ | + |
| | Seal *et al*. (2012) [22] | | - - | - - | + |
| | Shan *et al*. (2016) [23] | Trafic urbain | - - | ++ | + |
| | Tomaras *et al*. (2018) [25] | Maison intelligente et bâtiments | - - | + | + |
| | Spencer *et al*. (2018) [24] | | - - | - | - - |
| Réseaux Neuronaux (RN) | Kumar *et al*. (2013) [16] | Trafic urbain | - - | ++ | + |
| | Yu *et al*. (2005) [30] | Environnement | - - | - | - - |
| | Ma *et al*. (2020) [17] | | - - | ++ | + |
| | Pisa *et al*. (2019) [21] | | - - | ++ | - - |
| | Aliberti *et al*. (2018) [1] | Maison intelligente et bâtiments | - - | - - | - - |
| Régression+ RN | Zhu *et al*. (2015) [32] | Environnement | - - | ++ | + |
| | Mateo *et al*. (2013) [19] | Maison intelligente et bâtiments | - - | - | - - |
| Gradient Boost | Zhang et Haghani (2015) [31] | Trafic urbain | - - | - - | - - |
| Combiné | Oprea et Bâra (2019) [20] | Maison intelligente et bâtiments | - - | ++ | - - |

Comme le montre le tableau 2.1, la propriété d'ouverture n'est pas adressée par les techniques de l'état de l'art car l'ajout d'un grand nombre de capteurs en temps réel pourrait compromettre le fonctionnement du système, ou rendre des reconfigurations nécessaires. Nous pouvons également remarquer que la propriété d'hétérogénéité n'est que partiellement adressée car l'intégration d'informations dont le type n'est pas connu *a priori* nécessiterait une modification de la procédure d'estimation. Enfin, concernant le passage à l'échelle, les solutions proposées n'explicitent pas clairement si leur fonctionnement est correct ou pas lorsque le nombre de capteurs augmente.

Dans ce contexte, la contribution de ce travail peut être divisée en quatre parties : **(1)** la définition d'une approche pour estimer des valeurs environnementales manquantes dans des environnements à grande échelle, **(2)** la définition et **(3)** l'évaluation d'une nouvelle méthode géospatiale pour l'estimation de valeurs environnementales sur la ville de Toulouse, **(4)** les premières avancées sur le déploiement du système dans le cadre du GIS neOCampus.

## 3. LE SYSTÈME HYBRIDIOT

L'objectif de cette section est de présenter brièvement le système HybridIoT qui est une solution multi-agent pour estimer des valeurs environnementales dans des points non pourvus de capteurs physiques, tout en adressant les propriétés d'ouverture et d'hétérogénéité. L'utilisation du paradigme multi-agent permet (*i*) d'introduire de nouveaux agents de détection au moment de l'exécution sans reconfigurer la technique (ouverture) et (*ii*) d'estimer des informations hétérogènes sans configuration dépendant du type de données.

Le système HybridIoT se compose de deux types d'agents distincts : l'*Agent Capteur Réel* (*Real Sensor Agent*, RSA) qui représente toute instrumentation physique capable de fournir des informations précises sur l'environnement (comme la température, l'humidité, ...), et l'*Agent Contexte Ambiant* (*Ambient Context Agent*, ACA) qui est chargé d'estimer, en un point spécifique de l'environnement, la valeur qu'aurait perçue un capteur réel s'il était situé à ce même point. Un ACA peut être associé à un capteur physique ; son rôle est alors de fournir un mécanisme de résilience pour produire des estimations lorsque le capteur physique situé au même endroit est momentanément dans l'impossibilité de percevoir l'environnement (panne, ...).

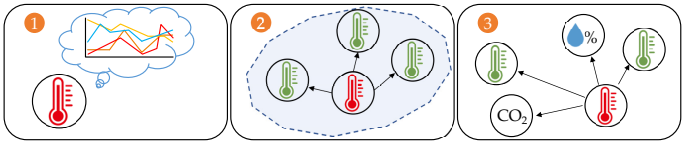



FIGURE 3.1 – Techniques d'estimation d'HybridIoT : estimation endogène à partir d'un historique (❶), estimation endogène à l'aide de capteurs voisins homogènes (❷) et estimation exogène à l'aide de données hétérogènes (❸).

HybridIoT propose d'estimer des informations environnementales selon trois mécanismes distincts (figure 3.1) :

(1) **Estimation endogène à partir d'un historique** ❶ : l'estimation se fonde sur les informations précédemment acquises par les capteurs ;
(2) **Estimation endogène à l'aide de capteurs voisins homogènes** ❷ : ce type d'estimation se fonde sur un critère de voisinage géospatial. Deux informations sont dites *homogènes* si elles sont de même type (par exemple température) et utilisent la même unité de mesure (degré Celsius). Ce type d'estimation utilise les informations de même type (et avec la même unité) acquises par les capteurs situés à proximité de l'endroit où on souhaite connaître une valeur environnementale ;
(3) **Estimation exogène** ❸ : l'estimation est réalisée en intégrant des informations hétérogènes (de types différents, avec des échelles différentes).

Chaque agent du système HybridIoT récupère avec une fréquence fixe, les informations en provenance de son capteur. Cette fréquence de récupération des informations captées est identique entre tous les agents du système.

Cet article présente deux de ces mécanismes : (*i*) la méthode d'estimation basée sur l'historique de données car elle introduit des notions fondamentales pour la compréhension de la suite du papier, et (*ii*) la méthode d'estimation endogène à l'aide de capteurs voisins homogènes. La méthode d'estimation exogène (*iii*) n'est pas présentée ici ; elle est consultable dans l'article de revue [8]. La méthode d'estimation endogène présentée ici est une nouvelle méthode d'estimation géospatiale basée sur l'utilisation d'une « *zone de confiance* ».

### 3.1. Estimation endogène à partir d'un historique

L'estimation endogène à partir d'un historique (figure 3.1 [❶]) s'effectue en deux étapes : (*i*) l'évaluation coopérative des poids à associer aux informations acquises par les dispositifs, puis (*ii*) l'estimation des données manquantes en utilisant les poids précédemment évalués. Avant de détailler chacune de ces étapes nous allons poser quelques définitions.

Définition 1 (**Fenêtre contextuelle**). — *Une* ACW *(Ambient Context Window)* $C_t$ *regroupe les informations environnementales homogènes (même type, même échelle) perçues par un dispositif durant un intervalle de temps discret* $T = [t - \delta, t],\ t - \delta < t$, *où* $t$ *est un indice qui correspond à la fin de l'intervalle de temps durant lequel l'information a été perçue. Une ACW contient* $|C_t| = |T|$ *informations homogènes. L'intervalle temporel entre les informations des ACW dépend de la fréquence d'acquisition des dispositifs.*

Définition 2 (**Distance entre ACW**). — *La distance entre deux ACW, calculée afin de les comparer, est définie comme la différence en valeur absolue des valeurs des deux ACW, divisée par le nombre de valeurs* $\gamma$ *des deux ACW. Plus la différence est faible, plus les deux ACW sont similaires. La distance entre deux ACW* $C_t$ *et* $C_k$ *est définie par la formule suivante :*

$$d\left(C_t, C_k\right) = \frac{\sum_{\ell \in [1,\gamma]} |E_\ell^t - E_\ell^k|}{\gamma} \tag{3.1}$$

*où* $\gamma = |C_t| = |C_k|$, $\ell$ *est un indice dans la plage* $[1, \gamma]$, $E_\ell^t$ *et* $E_\ell^k$ *sont respectivement les valeurs en position* $\ell$ *dans* $C_t$ *et* $C_k$.

Supposons qu'un *Ambient Context Agent*, $\text{ACA}_i$, doive estimer une information manquante à l'instant $t$ parce que son capteur associé n'est plus disponible. L'$\text{ACA}_i$ cherche parmi ses ACW (son historique), celles qui sont les plus similaires à celle contenant l'information à estimer. Plus concrètement, l'$\text{ACA}_i$ utilise les distances entre les ACW pour calculer un poids $w_t$ qui est ajouté à la dernière information perçue à l'instant $t - 1$.

Soit $\xi$ un sous-ensemble des ACW dans l'historique de l'$\text{ACA}_i$ tel que la distance $d(C_t, C_k)$ soit minimale, $\forall C_k \in \xi, k \neq t$ où $C_t \notin \xi$ est l'ACW qui contient l'information à estimer à l'instant $t$. La taille de $\xi$ (10) a été décidée de manière empirique. Le poids $w_t$ est calculé comme la moyenne des différences des deux dernières valeurs de chaque ACW $C_k \in \xi$ en utilisant la distance $d(C_k, C_t), \forall C_k \in \xi$. Le poids $w_t$ est calculé grâce à la formule (3.2) :

$$w_t = \frac{\sum_{C_k \in \xi} \left(E_\ell^k - E_{\ell-1}^k\right) \cdot \left(1 - d\left(C_t, C_k\right)\right)}{\sum_{C_k \in \xi} \left(1 - d\left(C_t, C_k\right)\right)} \tag{3.2}$$

où $C_t$ est l'ACW contenant l'information à estimer à l'instant $t$, $C_k \in \xi$ est la $k$-ième ACW la plus similaire à $C_t$ et $E_\ell^k$ et $E_{\ell-1}^k$ sont respectivement la $\ell^e$ et la $(\ell-1)^e$ valeur de l'ACW $C_k \in \xi$, à savoir les deux dernières entrées de contexte de $C_k$. La distance $d$

dans la formule (3.2) est normalisée dans l'intervalle $[0, 1]$ de manière à donner plus d'importance aux ACW qui sont moins éloignées (donc sont plus similaires) à l'ACW $C_t$.

L'estimation de la valeur manquante est alors calculée à l'aide de la formule :

$$E_t^j = E_{t-1}^j + w_t \tag{3.3}$$

où $E_t^j$ est l'information manquante à l'instant $t$, et $E_{t-1}^j \in C_j$ est la dernière information perçue par l'ACA$_i$.

Enfin, l'ACA$_i$ évalue une nouvelle ACW pour l'information estimée à l'instant $t$ [7].

Un processus de coopération entre ACA a été défini pour éviter d'estimer des informations à partir de données bruitées. Ce processus est initié lorsqu'un ACA$_i$ estime une information de manière endogène et lorsque d'autres agents percevant le même type de donnée que l'ACA$_i$ sont présents dans l'environnement. Dans une telle situation, l'ACA$_i$ choisit les agents avec qui il coopère selon deux critères : soit avec les agents les plus proches physiquement, soit avec les agents qui suivent une dynamique d'information similaire à la sienne. Deux agents (ACA) suivent une dynamique d'information similaire si les gradients de la fonction $f_i$ de chaque ACA (avec $f_i : T \rightarrow V$, ou $T \subset \mathbb{N}$ et $V \subset \mathbb{R}$, où $f_i(t)$ est la valeur perçue par l'ACA $i$ à l'instant $t$) sont similaires. Lorsque plusieurs agents suivent une dynamique similaire ou sont à une distance identique de l'ACA$_i$, ceux choisis sont ceux qui ont déjà fourni des valeurs utiles au processus d'estimation. L'utilisateur du système doit décider quel critère de coopération entres ACA il souhaite mettre en oeuvre (voisinage spatial ou dynamique d'information similaire) [6]. Ensuite, l'ACA$_i$ indique à ces agents quelles ACW de leur historique ils doivent utiliser pour calculer les poids qui seront utilisés par ACA$_i$, pour faire son estimation. Chaque agent fournit alors à l'ACA$_i$ un poids calculé à l'aide de la formule (3.2). L'ACA$_i$ calcule autant d'estimations que de poids reçus par les agents avec qui il coopère. Toutes les valeurs sont enfin pondérées afin de fournir une valeur d'estimation [8].

La figure 3.2 résume les étapes du processus de coopération entre les ACA.

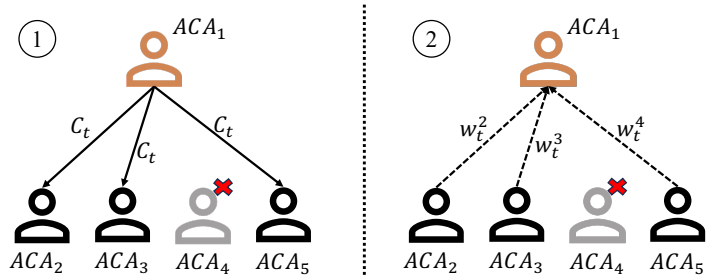



FIGURE 3.2 – Un ACA (en orange) coopère avec les agents (en noir) présents dans sa zone de confiance, afin d'estimer l'information manquante.

À l'étape ①, l'$ACA_1$ (devant estimer la valeur manquante à l'instant $t$) fournit aux autres agents présents dans sa zone de confiance la fenêtre contextuelle $C_t$. L'$ACA_4$ est grisé, car il n'est pas situé dans le voisinage de l'$ACA_1$ ; il ne participe donc pas au processus coopératif. Les agents coopératifs ($ACA_2$, $ACA_3$, $ACA_5$) comparent $C_t$ aux fenêtres contextuelles (ACW) présentes dans leur historique. Ils utilisent respectivement les fenêtres contextuelles les plus similaires (en utilisant la formule (3.1)) à $C_t$ dans leur historique pour calculer les poids d'estimation (en utilisant la formule (3.2)) qui sont ensuite utilisés par l'$ACA_1$ afin d'estimer l'information manquante (étape ②).

### 3.2. Estimation endogène à l'aide de capteurs voisins homogènes

Dans ce type d'estimation (figure 3.1 [❷]), l'*Ambient Context Agent*, $ACA_i$, estime les valeurs manquantes en coopérant avec des agents (soit des RSA (*Real Sensor Agent*), soit des ACA) percevant le même type d'information que lui et situés à l'intérieur de sa zone de confiance. Ce type d'estimation est fondé sur un critère de proximité géospatiale.

Définition 3 (**Zone de confiance**). — *Une zone de confiance est un maillage structuré de triangles équilatéraux de même taille (donnée) et associé à un ACA.*

La zone de confiance délimite une partie locale de l'environnement dans laquelle les capteurs fournissent des informations qui suivent une dynamique similaire (c'est-à-dire ayant des ACW dont la distance est faible). La forme de la zone de confiance peut être modifiée afin de ne garder dans la région que les capteurs qui fournissent des informations similaires aux siennes.

Quand un ACA est incapable de percevoir une information à partir de l'observation directe de son environnement, il coopère avec les agents situés dans sa zone de confiance. Ces derniers lui envoient leurs perceptions de l'environnement. Pour chaque paire d'agents dans la zone de confiance, l'$ACA_i$ calcule un unique **champ de données** représentant une estimation pour la valeur manquante.

Définition 4 (**Champ de données**). — *Un* champ de données $\Gamma$ *entre deux agents (soit ACAs soit des RSAs) est un champ vectoriel dans l'espace euclidien. Chaque point est associé à un vecteur qui est orienté vers l'agent qui fournit la valeur la plus élevée.*

L'estimation se base alors sur la valeur du gradient entre les données perçues par les capteurs. La figure 3.3 montre un exemple de champ de données généré par deux ACA.

Une fois les champs de données calculés, l'$ACA_i$ calcule l'estimation de la valeur manquante à l'instant $t$.

$$E_t^j = \frac{\sum_{p \in \text{keyset}(\Theta)} \Gamma(p) \cdot \Theta(p)}{\sum_{p \in \text{keyset}(\Theta)} \Theta(p)}. \tag{3.4}$$

où $E_t^j$ est l'estimation de l'information manquante à l'instant $t$, $\Gamma(p)$ est le champ de données calculé par la paire d'agents $p$, $\Theta$ est un dictionnaire contenant des paires

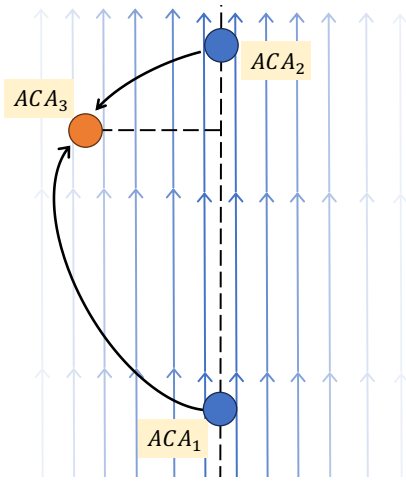



FIGURE 3.3 – Le champ de données généré par l'$ACA_1$ et l'$ACA_2$ est utilisé par l'$ACA_3$ pour effectuer une estimation. Les lignes en bleu représentent la direction du champ de données (elles sont orientées de la valeur la plus basse vers la plus haute). La transparence des flêches indique la précision du champ de données : plus l'$ACA_3$ est proche du champ reliant l'$ACA_1$ et l'$ACA_2$, plus le champ de données est précis.

d'agents comme clé et leur colinéarité par rapport à l'$ACA_i$ comme valeur. Trois points P1, P2, P3, sont dits colinéaires s'ils sont situés sur une même droite [6]. La valeur de colinéarité quantifie l'alignement des trois points. L'équation (3.4) permet de calculer une estimation de la valeur manquante en pondérant les champs de données des paires d'agents évaluées avec les valeurs de colinéarité des agents par rapport à $ACA_i$. L'idée est que les agents ayant la valeur de colinéarité la plus faible (c'est-à-dire étant les plus alignés) fournissent des informations plus précises car le gradient correspondant fournit des valeurs précises si l'$ACA_i$ est proche de la ligne qui relie l'$ACA_1$ et l'$ACA_2$.

keyset($\Theta$) contient les paires d'agents à l'intérieur de la zone de confiance, $\Theta(p)$ est la colinéarité entre les agents de la paire $p$ et l'$ACA_i$.

Enfin, la forme de la zone de confiance de l'$ACA_i$ est modifiée pour ne garder à l'intérieur de celle-ci que les agents qui ont fourni des valeurs cohérentes (notion définie par la suite).

Pour modifier la forme de la zone de confiance, l'$ACA_i$ trie la liste des valeurs des champs de données reçues selon l'ordre croissant puis calcule le champ de données médian $\widetilde{\gamma}$. L'$ACA_i$ calcule ensuite la valeur $\delta$ comme étant la distance entre les champs de données de la première paire d'agents ($\Gamma(p_1)$ pour $p_1$) et de la troisième ($\Gamma(p_3)$ pour $p_3$). Nous avons choisi la première et la troisième paires d'agents de manière empirique, après plusieurs expérimentations effectuées en environnements fermés (intérieurs), mais aussi en environnements ouverts (extérieurs) [10]. Nous utilisons deux valeurs

de seuil $th^{\pm}$ pour partitionner les champs de données.

$$th^{\pm} \leftarrow \Gamma(p_1) \pm \delta \cdot \omega \tag{3.5}$$

où la constante $\omega$ a une valeur fixée expérimentalement à $2,5$. Des valeurs sont dites *cohérentes* si elles appartiennent à l'intervalle $[th^{-}, th^{+}]$.

Au début de son fonctionnement, un ACA est associé à un maillage de taille limité qui ne peut pas être modifié : celui-ci garantit une zone de confiance de taille minimale à l'ACA.

Nous avons supposé que la zone de confiance contient des capteurs dont les valeurs perçues peuvent être utilisées pour décider s'il faut élargir ou pas la zone. Lorsqu'aucun capteur n'est disponible dans la zone de confiance, l'ACA élargit celle-ci dans toutes les directions. De cette manière, l'ACA peut explorer l'environnement afin de rechercher des capteurs qui peuvent contribuer à l'estimation de données manquantes.

L'algorithme de modification de la zone de confiance se base sur l'ajout ou l'élimination de triangles du maillage qui représente la zone de confiance de l'ACA.

La figure 3.4 montre un exemple d'exploration faite par l'ACA représenté par un carré bleu quand aucun capteur physique n'est disponible dans sa zone de confiance. L'exemple montre quatre étapes intermédiaires. Le processus d'agrandissement de la zone de confiance se poursuit jusqu'à que l'ACA rencontre au moins trois capteurs physiques. Les valeurs collectées par ces nouveaux capteurs physiques seront utilisées pour déterminer si la zone de confiance doit continuer à s'agrandir ou rétrécir (et vers quelle direction).

Dans l'exemple de la figure 3.4, un nouveau triangle est ajouté à chaque itération sur chaque côté du maillage. La zone de confiance croît dans toutes les directions jusqu'à ce qu'elle englobe au moins trois capteurs physiques pouvant fournir des données utiles au processus d'estimation (autrement dit des capteurs physiques percevant le même type d'information que l'ACA représenté par le carré bleu). L'ACA est alors capable de déterminer la direction vers laquelle il doit agrandir la zone. Il étend sa zone de confiance vers les capteurs dont les valeurs fournies sont dans la plage de valeurs déterminée par l'équation (3.5). Plus précisément, pour élargir sa zone de confiance en direction d'un capteur physique, l'ACA calcule d'abord la valeur de colinéarité entre sa position, celle du capteur physique et celles des centroïdes de tous les triangles situés sur le bord du maillage et dans la même direction que le capteur physique. Les valeurs de colinéarité sont ensuite normalisées dans l'intervalle [0, 100]. On ajoute peu à peu un triangle en direction des capteurs qui ont une valeur de colinéarité au dessus d'un seuil (fixée empiriquement à 80).

L'agrandissement de la zone de confiance cesse lorsqu'un capteur nouvellement inséré fournit des informations très éloignées de celles fournies par les capteurs inclus dans la zone. Il est également possible d'intégrer d'autres critères d'arrêt pour l'agrandissement de la zone de confiance, comme par exemple la distance maximale entre la bordure de la zone et l'ACA, une surface maximum, etc.

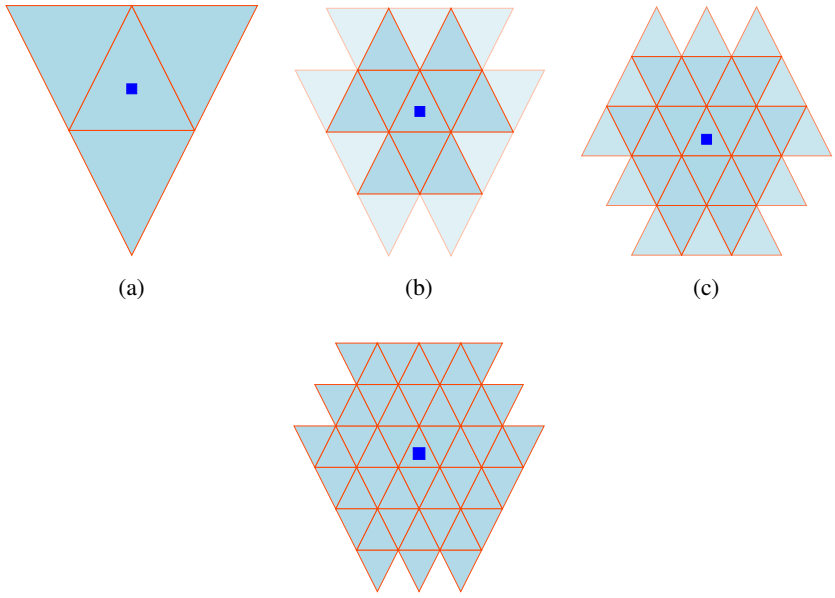



FIGURE 3.4 – Étapes intermédiaires de l'exploration de l'environnement par l'ACA (carré bleu) lorsqu'aucun capteur n'est disponible dans sa zone de confiance. La zone est agrandie par l'ACA selon les schémas de (a) à (d).

La figure 3.5 montre le processus de modification de la zone de confiance de l'ACA représenté par le carré bleu, lorsqu'un capteur physique (représenté par l'étoile violette) est présent dans cette zone de confiance et perçoit le même type d'information que l'ACA. Cet exemple montre les étapes intermédiaires du processus d'agrandissement dans toutes les directions du maillage représentant la zone de confiance.

L'opération de réduction de la zone de confiance fonctionne de manière opposée. Lorsqu'un capteur physique situé à l'intérieur de la zone de confiance fournit des valeurs qui sont très éloignées de celles perçues par les autres capteurs physiques situés dans la zone de confiance, le maillage doit être réduit pour peu à peu exclure ce capteur. Les principes utilisés sont les mêmes que précédemment. Plus précisément, pour diminuer sa zone de confiance en direction d'un capteur physique, l'ACA calcule d'abord la valeur de colinéarité entre sa position, celle du capteur physique et celles les centroïdes de tous les triangles situés sur le bord du maillage et dans la même direction que le capteur physique. Les valeurs de colinéarité sont ensuite normalisées dans l'intervalle [0, 100]. On enlève peu à peu un triangle en direction des capteurs qui ont une valeur de colinéarité au dessous d'un seuil (fixée empiriquement à 80).

L'utilisation d'un maillage pour représenter la zone de confiance présente deux avantages principaux : *(i)* la géométrie nécessaire pour agrandir ou rétrécir la zone de confiance, se limite à l'ajout de triangles aux triangles situés à la frontière de la zone ;

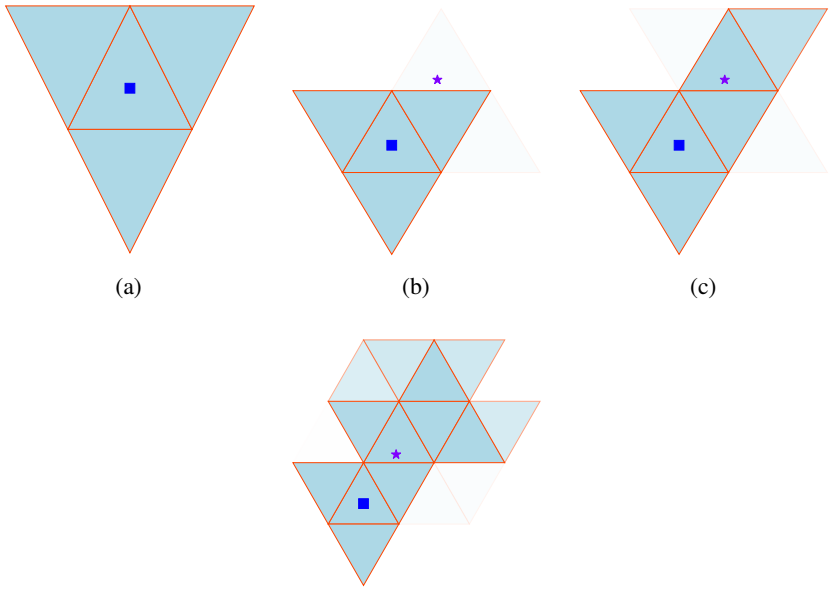



FIGURE 3.5 – Exemple de modification de la zone de confiance par l'ACA (carré bleu). Celle-ci s'étend dans la direction où le capteur physique (l'étoile) est situé. Dans cet exemple nous supposons que ce capteur fournit des valeurs pertinentes pour l'estimation des informations effectuée par l'ACA.

*(ii)* le seul paramètre requis pour gérer la zone de confiance est la longueur du côté des triangles. Il est important de souligner que l'utilisation d'un maillage présente également quelques limites. Un maillage peut entraîner des estimations biaisées, en particulier à proximité des murs des bâtiments, car les agents considèrent uniquement le critère de proximité spatiale entre les capteurs pour mettre en place un processus de coopération, et ce sans tenir compte d'éventuels obstacles pouvant avoir une influence sur la qualité des résultats. Par conséquent, le seul critère de proximité spatiale peut s'avérer insuffisant. Une solution serait d'associer aux données captées des informations environnementales précisant le contexte dans lequel est situé le capteur, afin de garantir des estimations plus précises.

## 4. Résultats expérimentaux

Cette section présente l'utilisation du système HybridIoT sur un jeu de données environnementales acquises à partir de capteurs déployés dans la ville de Toulouse. Après avoir décrit le scénario de simulation qui se base sur un ensemble de stations de météo situées dans la ville de Toulouse, nous décrivons le type de données utilisées, les expérimentations ainsi que les métriques utilisées pour évaluer les estimations fournies par HybridIoT. Nous présentons ensuite les résultats obtenus par HybridIoT lorsqu'il

utilise l'estimation par historique de données, puis lorsqu'il utilise l'estimation par zone de confiance. Nous comparons enfin les résultats obtenus avec des techniques d'estimation issues de l'état de l'art.

## 4.1. Description du Scénario

Les données utilisées, issues des stations météo installées sur la ville de Toulouse, sont disponibles en libre accès sur le site Web de Toulouse Metropole[(1)]. Nos expérimentations se focalisent sur la température (en degrés Celsius). Les données de températures ont été collectées pendant 11 jours, du 27 juin 2020 au 8 juillet 2020, avec une fréquence d'acquisition de 15 minutes. Les jours où certains capteurs n'étaient pas opérationnels ne sont pas pris en compte. Plus concrètement, cet ensemble de données correspond à un tableau de $1\,000 \times 19$ valeurs numériques.

La figure 4.1 montre la position des capteurs à partir desquels les données utilisées sont issues.

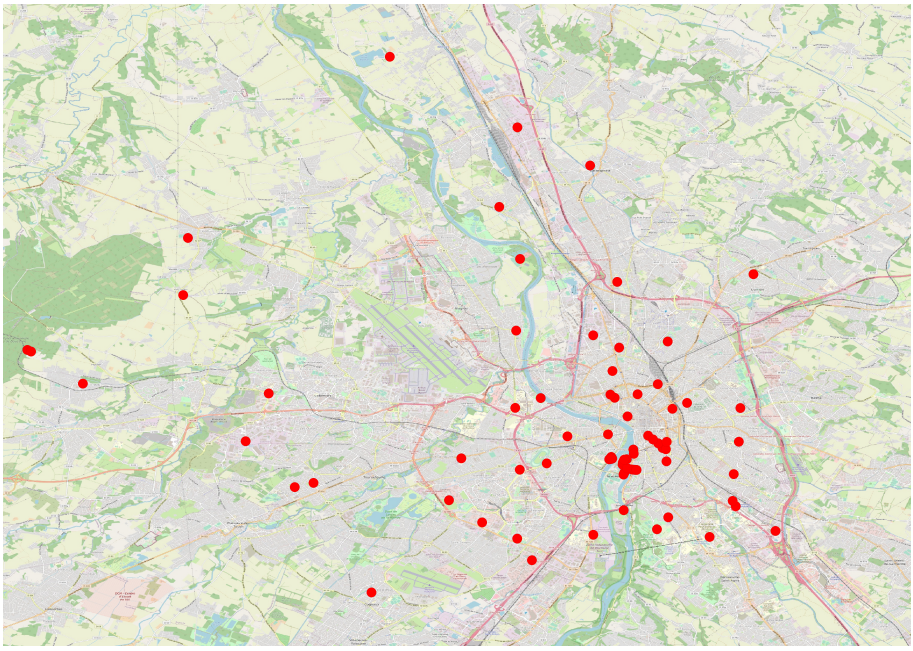


Figure 4.1 – Position des capteurs (stations météo) disponibles dans la ville de Toulouse.

[(1)] `https://data.toulouse-metropole.fr/explore/?sort=modified`; Dernière visite : 20 Juillet, 2023

### 4.2. Évaluation

Pour évaluer la précision des résultats obtenus à l'aide de nos estimations, nous utilisons les mesures d'erreur suivantes :

- **Erreur absolue d'estimation (MAE)** : l'erreur absolue moyenne est calculée comme la différence entre l'estimation et la valeur réelle. Cette valeur représente la grandeur de l'erreur. Cette mesure est exprimée dans la même unité que l'information traitée. Plus la valeur est petite, meilleur est le résultat.
- **Pourcentage d'erreur d'estimation (MAPE)** : le pourcentage d'erreur absolue moyenne est calculé comme la différence entre l'estimation et la densité réelle, divisée par la valeur réelle. Cette valeur représente le pourcentage d'écart entre les valeurs estimées et réelles. Plus le pourcentage est petit, meilleur est le résultat.

Nous utilisons la validation croisée de séries temporelles pour évaluer la précision des résultats de l'estimation. Soit $y = y_1, \ldots, y_n$ une série temporelle. Dans la validation croisée de séries temporelles, les échantillons de $y_1$ à $y_{t-1}$, avec $1 \leqslant t - 1 < n$, sont utilisés comme ensemble d'apprentissage pour l'estimation de $y_t$. Contrairement à la validation croisée standard, à chaque instant $t$ nous n'utilisons que les valeurs antérieures à $t$, de l'instant 0 à l'instant $t - 1$.

Nous avons exécuté chaque simulation 99 fois pour chaque configuration. Une configuration comprend deux paramètres :

- le pourcentage de capteurs manquants (de 10 % à 90 %, par incrément de 10 %) : cette valeur indique combien de capteurs disponibles seront remplacés par des ACA. La permutation des capteurs remplacés par des ACA est choisie de manière aléatoire ;
- le pourcentage de données manquantes pour chaque capteur (de 10 % à 90 %, par incrément de 10 %) : nous supposons que des capteurs virtuels peuvent être associés à des capteurs physiques qui ne peuvent pas être toujours disponibles (à cause de pannes ou d'opérations de maintenance). Pour simuler l'intermittence des capteurs, nous estimons juste un certain pourcentage de données manquantes.

Pour chaque combinaison des paramètres précédents, nous exécutons la méthode d'estimation 99 fois. Cela donne un nombre de simulations égal à 8 019.

Pour vérifier si 99 simulations sont suffisantes pour obtenir des résultats statistiquement significatifs, nous calculons la moyenne cumulée de l'erreur d'estimation pour toutes les simulations et pour chaque ACA. La figure 4.2 montre la moyenne cumulée de l'erreur d'estimation (MAE) obtenue pour chaque pourcentage de capteurs manquants. Pour chaque pourcentage, nous calculons la moyenne des erreurs d'estimation obtenues pour chaque combinaison de paramètres (pourcentage de capteurs manquants, pourcentage de données manquantes, identifiant ID de la simulation). Comme le montre la figure 4.2, l'erreur d'estimation moyenne cumulée converge vers une moyenne stable, ce qui montre que le nombre de simulations choisi est suffisant.

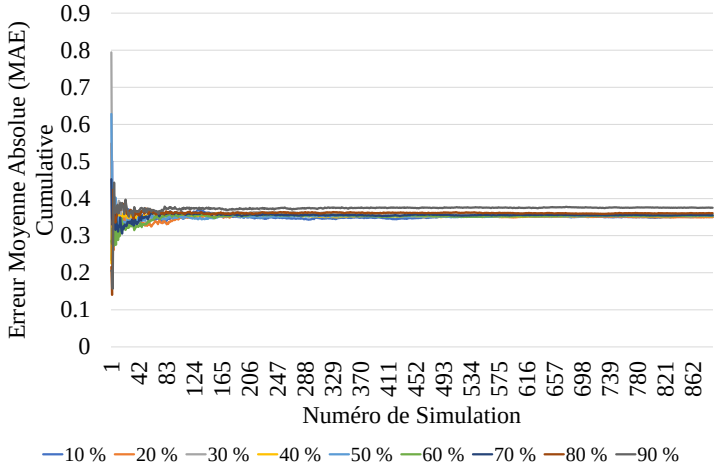



FIGURE 4.2 – Erreur moyenne absolue (MAE) cumulative. Le diagramme montre la stabilité de l'erreur moyenne des simulations faites.

## 4.3. ÉVALUATION DE L'ESTIMATION PAR HISTORIQUE DE DONNÉES

Cette section présente les résultats obtenus pour l'estimation endogène à partir de l'historique de données.

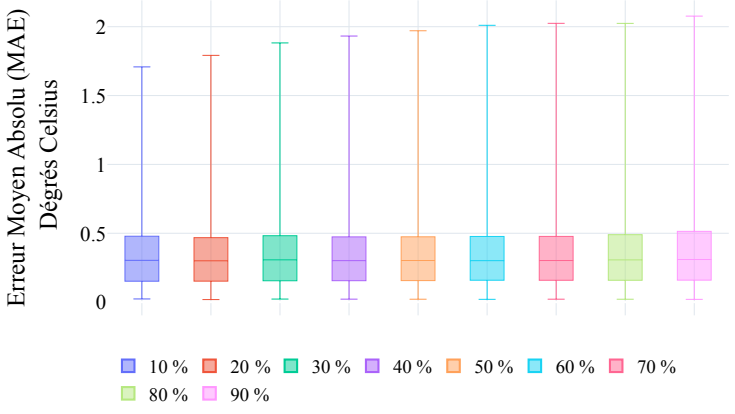



FIGURE 4.3 – Erreur moyenne absolue (MAE) obtenue par l'estimation par historique de données en fonction des différents pourcentages de capteurs manquants.

La figure 4.3 montre l'erreur absolue d'estimation (MAE) obtenue en utilisant l'historique de données. L'erreur MAE est obtenue en calculant, pour chaque capteur, la moyenne de l'erreur obtenue après 99 simulations. À chaque simulation, les capteurs

virtuels permutés sont choisis aléatoirement. Nous obtenons un erreur moyenne de 0,39 degrés Celsius et un écart moyen de 0,45 degrés. Au fur et à mesure que les capteurs disponibles diminuent, l'incertitude sur les estimations fournies augmente, car les informations peuvent être collectées depuis des parties de l'environnement qui suivent des dynamiques très différentes. Cependant, nous pouvons remarquer que l'erreur médiane est presque constante ce qui montre que la solution proposée par HybridIoT fournit des estimations précises.

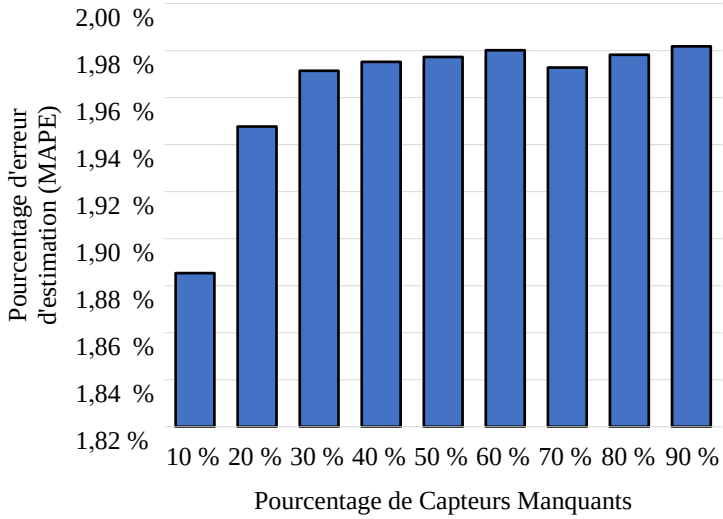



FIGURE 4.4 – Erreur Moyenne Absolue en Pourcentage (MAPE) en utilisant l'historique des données.

La figure 4.4 montre l'erreur moyenne absolue en pourcentage (MAPE) obtenue en utilisant l'historique de données, après 99 simulations. Nous pouvons remarquer que le pourcentage d'erreur, compris entre 1,88 % et 1,98 %, croît au fur et à mesure que le nombre des capteurs physiques disponibles diminue.

La figure 4.5 montre le diagramme de Pareto obtenu à partir de l'erreur MAE. Ce type de diagramme donne une indication sur la distribution des erreurs obtenues pour toutes les simulations faites en utilisant l'historique de données.

Dans ce diagramme, les histogrammes indiquent la fréquence des erreurs dans les simulations (axe des ordonnées de gauche). Ces mesures d'erreur ont été pondérées par la distance de l'erreur moyenne calculée parmi toutes les simulations. La courbe indique le pourcentage cumulé du nombre total d'occurrences des erreurs (axe des ordonnées de droite). La majorité des erreurs obtenues par les techniques proposées sont situées dans les histogrammes les plus à gauche, qui représentent plus du 65 % des simulations réalisées. En outre, moins de 30 % des simulations ont produit des valeurs estimées éloignées de plus de 0,23 degrés Celsius de l'erreur moyenne.

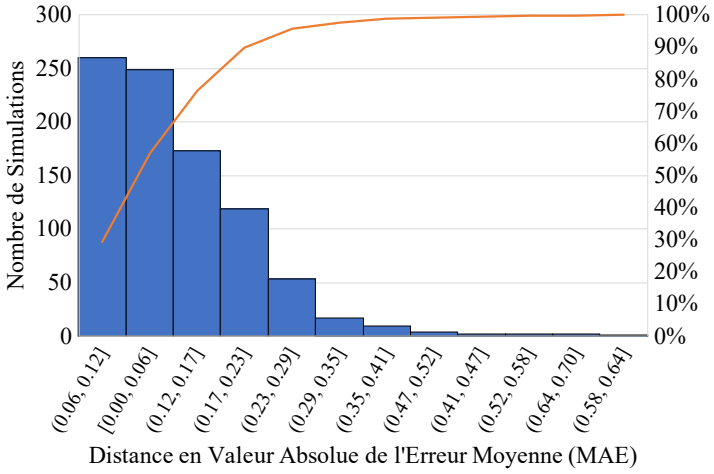



FIGURE 4.5 – Diagramme de Pareto obtenu à partir de la distance absolue entre les erreurs moyennes (MAE) des simulations individuelles, et l'erreur moyenne (MAE) de toutes les simulations. Les erreurs sont obtenues en utilisant l'historique de données.

### 4.4. ÉVALUATION DE L'ESTIMATION PAR ZONE DE CONFIANCE

L'estimation par zone de confiance est utilisée lorsque l'ACA ($ACA_i$) souhaitant faire une estimation, possède, dans sa zone de confiance, des capteurs physiques qui perçoivent des informations de même type et de même unité que celles qu'il perçoit.

L'utilisation de cette technique nécessite qu'au moins 3 capteurs physiques soient disponibles dans la zone de confiance de l'$ACA_i$. À l'initialisation, l'$ACA_i$ est associé à une zone de confiance de taille limitée (un triangle). La phase d'exploration de l'environnement par l'$ACA_i$ peut permettre d'atteindre des parties de l'environnement contenant des capteurs physiques pouvant contribuer à l'estimation de données manquantes. Si malgré cela, l'$ACA_i$ ne possède pas au moins 3 capteurs physiques dans sa zone de confiance, il estime l'information manquante en utilisant son historique de données. Les résultats présentés dans cette section ne sont donc pas exclusivement obtenus par zone de confiance, mais plutôt à l'aide de celle-ci.

La figure 4.6 montre l'erreur (MAE) obtenue par l'estimation par la zone de confiance, après 99 simulations.

Comme pour le cas précédent (estimation par historique), l'absence de capteurs entraîne une augmentation de l'erreur d'estimation. Néanmoins, la médiane de l'erreur est de moins de 0,5 degrés Celsius.

La figure 4.7 montre l'erreur moyenne absolue en pourcentage (MAPE) obtenue en utilisant la zone de confiance, après 99 simulations. Nous pouvons remarquer que le pourcentage d'erreur, compris entre 1,70 % et 1,87 %, croît au fur et à mesure

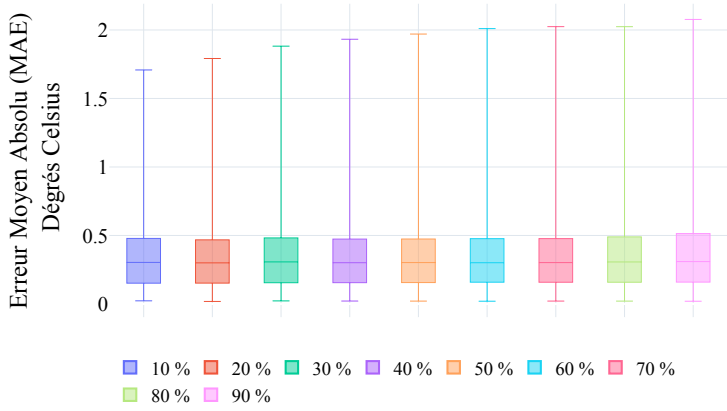



FIGURE 4.6 – Erreur moyenne absolue (MAE) obtenue par l'estimation par zone de confiance.

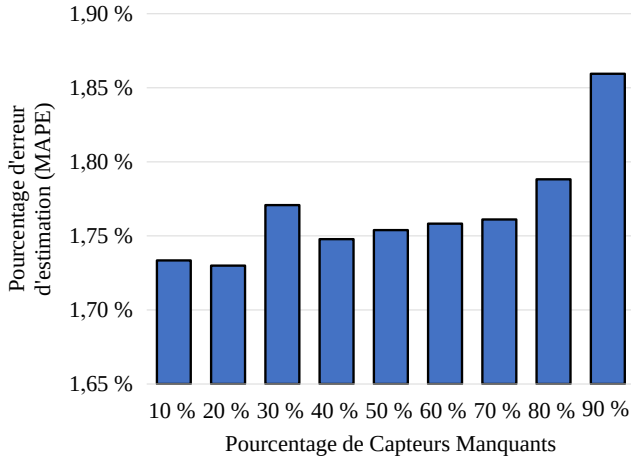



FIGURE 4.7 – Pourcentage d'erreur (MAPE) obtenu par l'estimation par zone de confiance.

que le nombre des capteurs physiques disponibles diminue. Les valeurs obtenues montrent que la technique d'estimation avec la zone de confiance permet d'obtenir des estimations précises.

La figure 4.8 montre le diagramme de Pareto obtenu à partir des erreurs MAE calculées par zone de confiance. La majorité des erreurs obtenues par les techniques proposées sont situées dans les histogrammes les plus à gauche, qui constituent plus de 70 % de toutes les simulations. De plus, moins de 30 % des simulations ont produit des valeurs estimées éloignées de plus de 0,12 degrés Celsius de l'erreur moyenne.

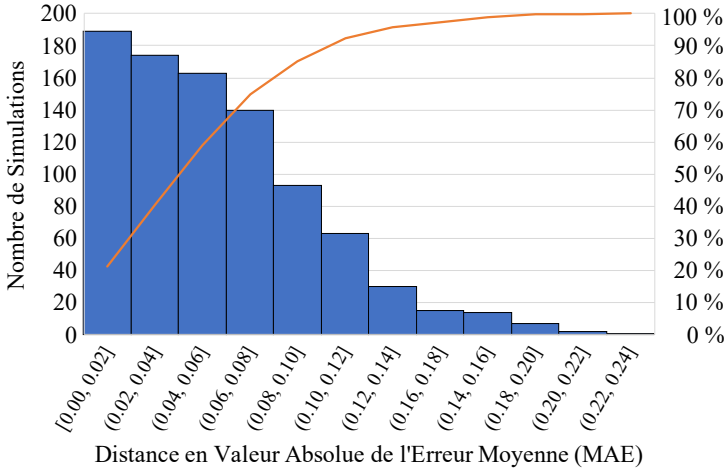
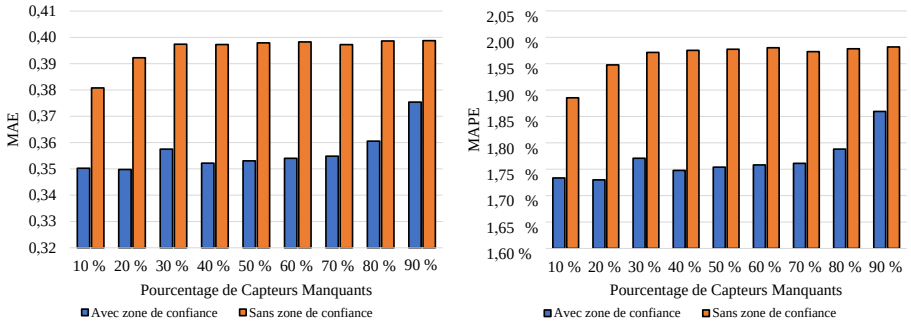



FIGURE 4.8 – Diagramme de Pareto obtenu à partir de la distance absolue entre les erreurs moyennes (MAE) des simulations individuelles, et l'erreur moyenne (MAE) de toutes les simulations. Les erreurs sont obtenues en utilisant la zone de confiance.

La figure 4.9 montre les erreurs MAE et MAPE obtenues en utilisant respectivement la zone de confiance combinée à l'historique de données, et l'historique de données uniquement.




(a) (b)

FIGURE 4.9 – Erreurs MAE (figure 4.9a) et MAPE (figure 4.9b) obtenues en utilisant la zone de confiance et l'historique de données.

Les résultats obtenus montrent que l'estimation par zone de confiance, éventuellement combinée à la technique par historique lorsque cela est nécessaire, permet d'obtenir des estimations précises. Ceci se justifie par le fait que l'estimation par zone de confiance estime des données à partir de capteurs déjà présents dans l'environnement qui fournissent des information perçus depuis l'environnement réel. De plus, les

estimations sont calculées à partir d'informations provenant de capteurs qui ne fournissent que des valeurs qui suivent une dynamique similaire dans des régions locales de l'environnement.

La figure 4.10 montre des exemples de zones de confiance obtenues pour des ACAs à la fin d'une simulation.

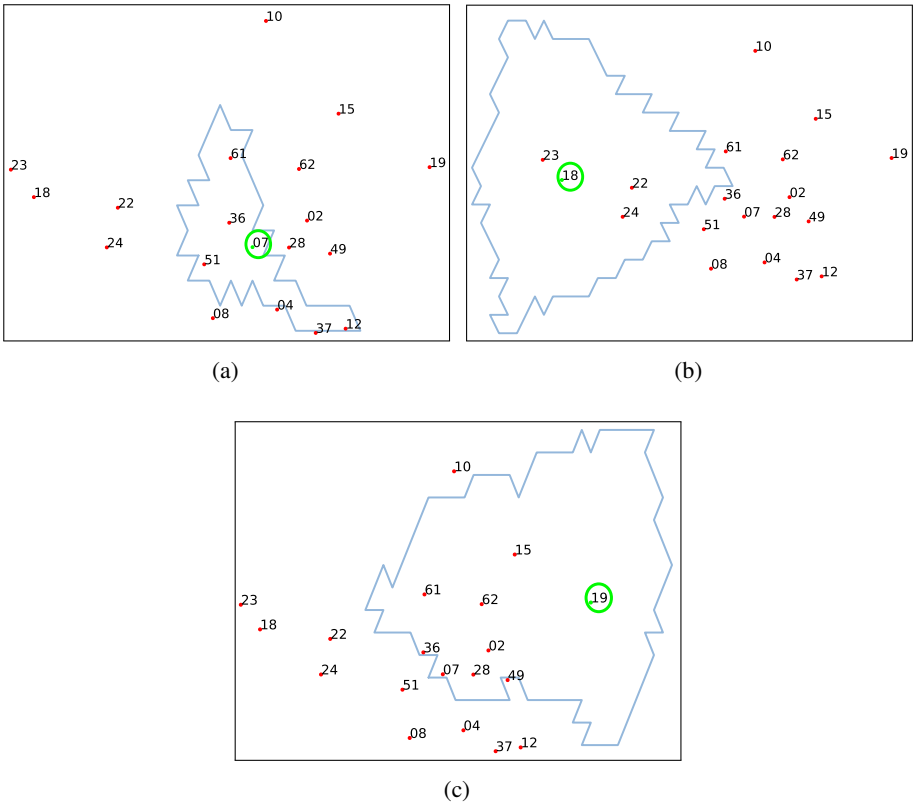



FIGURE 4.10 – Zones de confiance obtenues à la fin d'une simulation. Les points en rouge représentent les capteurs physiques disponibles, en verts les ACA.

Dans la figure 4.10a, la zone ne s'étend ni vers l'est ni vers l'ouest parce que les valeurs perçues par les autres capteurs sont différentes de celles acquises par les capteurs proches de l'$ACA_{07}$. Le nombre de capteurs disponibles autour de l'ACA permet d'adapter de manière précise la forme de la zone de confiance.

En revanche, dans les figures 4.10b et 4.10c on peut remarquer que les zones de confiance se sont agrandies lors de l'exploration de l'environnement par l'ACA qui est à la recherche de capteurs physiques qui pourraient lui être utiles pour son processus d'estimation.

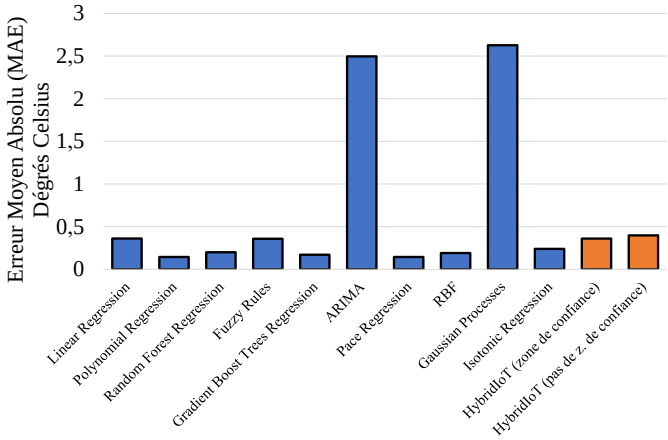
### 4.5. Comparaison avec des méthodes d'estimation issues de l'état de l'art

Les résultats obtenus par l'estimation endogène ont été comparés à ceux obtenus par des techniques de régression standard disponibles dans le logiciel KNIME. Ce logiciel gratuit offre un environnement graphique de type « glisser-déposer » dans lequel des *workflow* peuvent être assemblés en connectant des nœuds qui effectuent des tâches d'analyse de données [2]. Dans KNIME, les nœuds sont des composants représentés par des boîtes dotées de ports d'entrée et de sortie. Chaque nœud transforme et traite les données selon des fonctionnalités spécifiques. Les ports de connexion d'entrée/sortie permettent aux données de circuler dans le pipeline. La plateforme KNIME a été choisie pour cette évaluation en raison de sa disponibilité et de sa facilité d'utilisation. En outre, KNIME propose un grand nombre de techniques de régression, ce qui permet une comparaison plus exhaustive avec notre proposition. Les techniques de régression suivantes ont été utilisées pour estimer les informations manquantes : régression linéaire, régression polynomiale, régression par forêt aléatoire [14], règles floues [3], régression par arbres à gradient boost [4], moyenne mobile intégrée autorégressive (ARIMA) [15], régression Pace [27], fonction de base radiale (RBF) [26] et régression isotonique [29]. Les nœuds de régression, des techniques RBF et de régression isotonique sont disponibles dans la plate-forme d'algorithmes de data mining Weka [28], qui peut être intégrée à KNIME. Le tableau 4.1 résume les paramètres par défaut utilisés pour configurer les nœuds de régression dans KNIME.

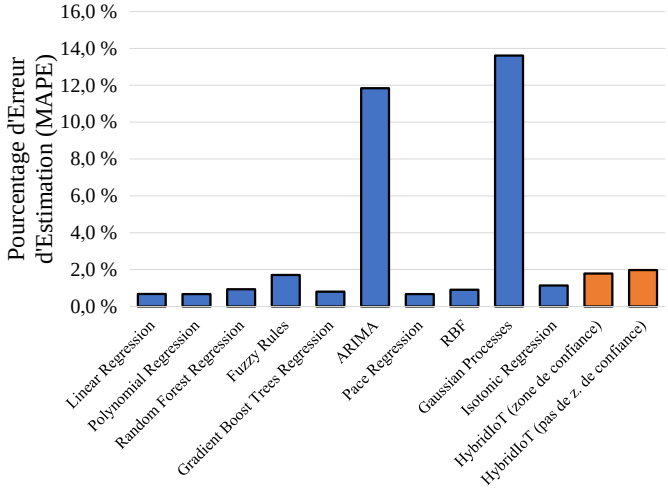
Table 4.1 – Configuration des techniques présentes dans l'état de l'art de KNIME

| **Technique** | **Propriétés** |
|---|---|
| Régression linéaire | Aucune propriété |
| Régression polynomiale | Degré polynomial maximum : 3 |
| Régression de forêt aléatoire | Aucune propriété |
| Règles floues | Valeurs manquantes : Meilleure estimation |
| Gradient Boost Trees Régression | Gestion des valeurs manquantes : XGBoost<br>Alpha : 1.0 |
| ARIMA | AR/I/MA : 1<br>Méthode d'estimation : vraisemblance conditionnelle [11] |
| Régression Pace | Estimateur : moindres carrés ordinaires<br>Valeur seuil : 2 |
| RBF | Nombre de fonctions de base gaussiennes : 2<br>Facteur de crête pour la pénalité quadratique sur les poids de sortie : 0,01<br>Paramètre de tolérance pour les valeurs delta : 1.0e-6<br>Option d'optimisation de l'échelle : une échelle par unité<br>Utiliser la descente de gradient conjugué : vrai<br>Utiliser les fonctions de base normalisées : vrai<br>Taille du pool de threads : 1<br>Nombre de threads à utiliser : 1<br>Utiliser la valeur de départ du nombre aléatoire : vrai |
| Processus gaussiens | Niveau de bruit gaussien par rapport à la cible transformée : 1<br>Noyau utilisé : polynôme |
| Régression isotonique | Aucune propriété |

La figure 4.11 compare l'erreur obtenue avec HybridIoT (avec et sans zone de confiance) par rapport aux techniques issues de l'état de l'art.




(a)




(b)

FIGURE 4.11 – Comparaison des erreurs MAE (figure 4.11a) et MAPE (figure 4.11b) obtenues par HybridIoT et les techniques de l'état de l'art disponibles dans le logiciel KNIME.

Globalement, les résultats montrent qu'HybridIoT estime de manière précise les informations manquantes. Par rapport aux techniques de l'état de l'art, nous pouvons également remarquer :

- qu'HybridIoT ne nécessite pas de configuration *a priori*.
- qu'HybridIoT considère pas seulement les données, mais aussi leur distribution spatiale.

- qu'HybridIoT fournit en continu un processus d'estimation et s'adapte au contexte environnemental. Autrement dit, les ACA sont toujours capables de fournir des estimations, même si des capteurs réels sont déplacés où sont momentanément indisponibles.

## 5. Déploiement dans un contexte réel

Cette section présente les activités en cours de réalisation dans le cadre du projet HybridIoT. Le projet a obtenu un financement de la part de Toulouse Tech Transfer pour être déployé au sein de neOCampus sur le campus de l'Université de Toulouse III - Paul Sabatier (UT3) afin d'atteindre un niveau de maturité technologique (TRL) de 6-7. Le développement d'un prototype réel d'HybridIoT sur le campus de l'UT3 actuellement en cours (fin 2023), réalisé avec la participation d'AKKODIS-France.

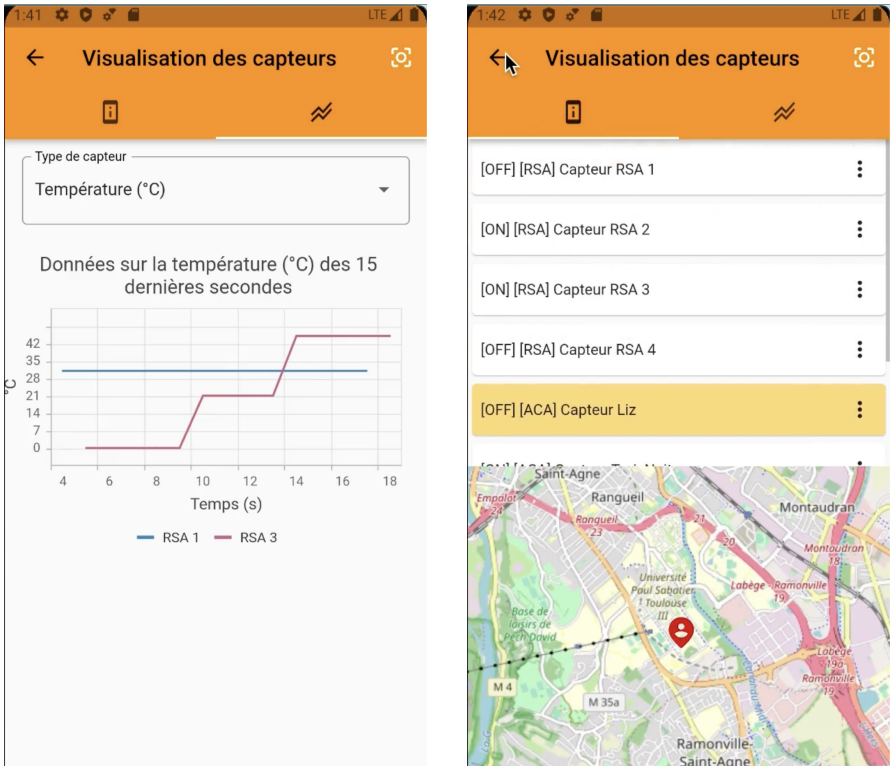
Le campus de l'UT3 s'étend sur environ 150 hectares et contient plus de 407 000 $m^2$ de bâtiments, avec 36 000 utilisateurs qui le fréquentent. Dans neOCampus, le campus est considéré comme une ville intelligente où plusieurs milliers de données sont fournies par des capteurs hétérogènes placés à l'intérieur et à l'extérieur des bâtiments ($CO_2$, consommation d'énergie et de fluides, humidité, luminosité, présence humaine, ...). Il s'agit donc, dans ce cadre, d'utiliser des capteurs IoT réels possédés par les utilisateurs du campus (RSA) ainsi que des capteurs virtuels (ACA).

Le développement front-end d'HybridIoT consiste à développer une application mobile multiplateforme, fonctionnant sous Android et IOS, qui joue le rôle d'intermédiaire entre l'estimation et les capteurs embarqués (si disponibles). Le développement back-end d'HybridIoT concerne l'infrastructure nécessaire (matérielle et logicielle) pour supporter le calcul de l'estimation de manière décentralisée ainsi que le stockage des informations acquises par les capteurs (physiques ou virtuels).

L'application mobile propose trois fonctionnalités : l'acquisition d'informations à partir des capteurs embarqués, l'affichage de statistiques sur les informations perçues et l'installation de nouveaux capteurs ACA virtuels. Les deux premières fonctionnalités ne sont accessibles qu'aux utilisateurs, tandis que l'installation des capteurs ACA virtuels est réservée au personnel technique du campus.

La figure 5.1 montre deux captures d'écran de l'application développée sous Android. Actuellement, l'application permet la visualisation de l'emplacement des capteurs sur le campus ainsi que les valeurs environnementales perçues par les capteurs physiques sous forme de graphique.

Du point de vue du back-end, nous devons avoir une infrastructure pour supporter HybridIoT mais aussi fournir des informations instantanées à tout moment et quel que soit l'endroit. Cette infrastructure est supportée par la plateforme neOCampus de l'UT3 qui comporte trois équipements principaux : (*i*) 16 cœurs, 64 Go de RAM et 2,7 To de stockage haut débit, (*ii*) un serveur de stockage de 45 To partagé avec la plateforme CloudMIP, et (*iii*) un réseau IoT dédié, filaire et sans fil.




(a) (b)

FIGURE 5.1 – HybridIoT pour Android : statistiques des valeurs de température sur les 15 dernières secondes figure 5.1a), emplacement des capteurs et leur identifiant (figure 5.1b).

Le serveur neOCampus comprend un broker MQTT qui permet à tous les capteurs ou actionneurs ou applications ambiant(e)s d'être interconnecté(e)s. Les capteurs/actionneurs sont reliés au réseau IoT par des dispositifs tels que Raspberry Pi et ESP8266. Ces appareils sont authentifiés par un mécanisme hébergé sur le serveur, qui inclut l'application sensOCampus pour gérer les capteurs ainsi qu'une base de données spécifique. Un réseau LoRaWAN sans fil est également déployé à l'échelle du campus afin de permettre aux dispositifs LoRa d'envoyer les données des capteurs dans un rayon de plus de 2 km. Les données sont ensuite acheminées vers le serveur principal du campus [5].

L'objectif de cette nouvelle version d'HybridIoT est triple :

- limiter la portée opérationnelle d'HybridIoT au campus universitaire. Afin d'éviter des résultats biaisés causés par des appareils qui ne sont pas situés sur le campus, seuls les capteurs mobiles se trouvant à l'intérieur du campus doivent envoyer des informations au serveur.

- permettre l'utilisation de données acquises par des capteurs pouvant supporter différentes fréquences d'acquisition. Le protocole actuel doit être modifié car les données sont échangées entre les capteurs et le serveur via le protocole MQTT.
- proposer un calcul décentralisé *embarqué*, afin de réduire la charge de calcul du serveur et permettre à un utilisateur de réaliser l'estimation directement depuis son smartphone plutôt que sur le serveur.

Le fonctionnement d'HybridIoT peut alors se résumer ainsi : les agents (RSA et ACA) acquièrent des données (à partir des capteurs physiques disponibles) qui sont envoyées au serveur neOCampus via MQTT. Le serveur stocke ces informations, estime les valeurs manquantes pour les ACA enregistrés, puis renvoie l'estimation à l'utilisateur. Les experts du domaine (utilisateurs du système) doivent alors déterminer la qualité des estimations par rapport aux valeurs réelles, afin de décider si l'installation d'un nouveau capteur physique est nécessaire ou pas ; si tel n'est pas la cas, les ACA sont utilisés. Cette qualité peut être jugée en comparant les estimations et les valeurs perçues par les capteurs physiques voisins où par un capteur mobile temporairement installé sur place.

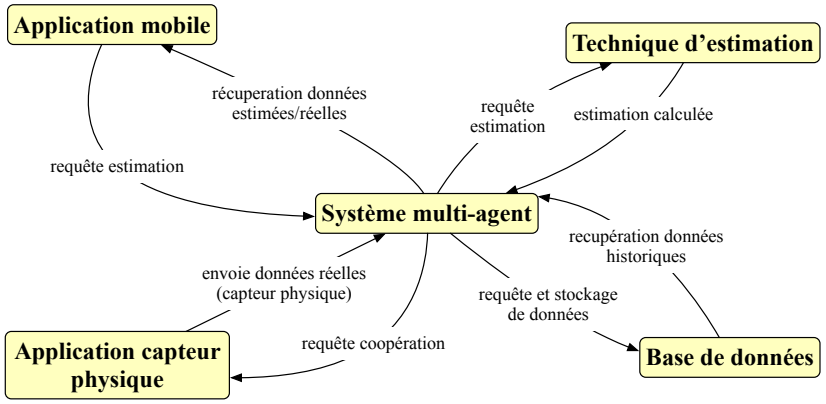



FIGURE 5.2 – Architecture du prototype du système HybridIoT.

La figure 5.2 montre l'architecture générale du prototype HybridIoT à déployer dans le campus UT3. Elle se compose de 5 composants :

- **Une base de données** qui permet de stocker les information estimées et réelles.
- **Une application mobile** qui permet d'envoyer une requête d'estimation pour estimer une valeur environnementale à un point d'intérêt, s'il n'existe pas de capteur réel au niveau de ce point d'intérêt.
- **Une application « capteur physique »** qui est associée à un capteur physique et peut envoyer les informations captées depuis l'environnement réel pour les utiliser dans le processus d'estimation.

- **Un système multi-agent** qui est le point central du système HybridIoT. Il détermine quelle technique d'estimation doit être lancée selon le contexte environnemental (Figure 3.1) et gère les interactions entre les agents ACA et RSA. Il est en particulier responsable de l'envoi des informations (estimées/réelles) aux agents, de la récupération des informations depuis la base de données puis de la transmission de ces informations à la technique d'estimation pertinente et enfin de l'envoi du résultat de l'estimation à l'application mobile et à la base de données.
- **Un ensemble de techniques d'estimation** qui correspond aux 3 techniques d'estimation définies dans HybridIoT. Chaque résultat de l'estimation est rajouté à la base de données pour pouvoir être ultérieurement utilisé, notamment lors de l'estimation par historique.

Lors de la mise en place de l'architecture d'HybridIoT sur le serveur neOCampus, nous avons utilisé Docker pour la « conteneurisation » des techniques d'estimation, pour faciliter le déploiement et la gestion des dépendances. À ce jour (fin 2023), le déploiement du prototype réel dans le campus est en phase d'essai.

## 6. Conclusion et travaux futurs

Cet article présente le système HybridIoT qui permet d'estimer des valeurs environnementales dans des endroits non couverts par des capteurs. Cette présentation se compose de trois parties complémentaires : (*i*) une méthode d'estimation basée sur l'utilisation conjointe de plusieurs techniques standards, à des fins de comparaison, (*ii*) une méthode d'estimation à base d'agents pour adresser simultanément les propriétés d'ouverture, d'hétérogénéité et de passage à l'échelle et (*iii*) le déploiement d'HybridIoT sur le campus de l'UT3. Cette dernière étape, en cours de réalisation (fin 2023), s'appuie sur l'infrastructure neOCampus, fournissant un serveur puissant pour stocker les données et effectuer les calculs. Elle est réalisée en collaboration avec l'entreprise AKKODIS, impliquée dans le développement du prototype dans le cadre d'un mécénat avec l'UT3.

HybridIoT a été évalué sur un jeu de données météorologiques et comparé avec des techniques de l'état de l'art. Les résultats obtenus montrent la pertinence de notre proposition pour estimer des valeurs environnementales manquantes dans des environnements à grande échelle utilisant un nombre limité de capteurs. Comparée aux techniques de l'état de l'art, HybridIoT permet d'adresser simultanément les propriétés d'ouverture, d'hétérogénéité et de passage à l'échelle. Plus de détails sur pour la pertinence des méthodes et les résultats sont disponibles dans [6].

Dans ce travail, nous avons utilisé le système d'HybridIoT dans le contexte de la ville intelligente. HybridIoT a également été utilisé pour déterminer des informations météorologiques (température, humidité, vitesse du vent, etc.) à l'échelle d'une région [8]. Il a également été utilisé pour estimer la densité du trafic urbain [9] à l'échelle d'une ville. Dans ce dernier cas, les agents utilisent les informations envoyées par les voitures connectées pour estimer la densité du trafic dans des parties locales d'un

environnement urbain. La diversité de ces applications montre la généricité de l'approche HybridIoT pour estimer des informations manquantes dans des environnements à échelles variables.

Nos objectifs à court terme sont *(i)* d'évaluer combien de capteurs physiques peuvent être remplacés par des ACA sans compromettre la qualité des estimations, *(ii)* de déterminer si l'installation de capteurs physiques est nécessaire dans les zones de l'environnement où les estimations ne sont pas cohérentes avec les valeurs réelles, et *(iii)* de déterminer l'influence de la topologie de l'environnement physique sur les estimations. Nous allons également poursuivre le déploiement d'HybridIoT sur le campus de l'UT3 afin d'atteindre un niveau de maturité technologique (TRL) de 6-7.

## 7. Remerciements

Nous remercions Nathan Stievano (étudiant alternant en M2 à l'UT3) et TTT qui finance cette alternance, et Mathieu Renard (étudiant stagiaire en M1 à l'UT3) pour le développement du prototype HybridIoT à déployer sur le campus UT3 (docker, développement base de données, déploiement dans le serveur neOCampus, mise en place des endpoint MQTT, développement de la nouvelle méthode d'estimation géospatiale) ainsi que neOCampus pour la labellisation du projet et l'accès à ses ressources déployées sur le campus. Nous remercions également AKKODIS pour son support au développement du prototype HybridIoT pour le campus UT3.

## Bibliographie

[1] A. Aliberti, F. M. Ugliotti, L. Bottaccioli, G. Cirrincione, A. Osello, E. Macii, E. Patti & A. Acquaviva, « Indoor Air-Temperature Forecast for Energy-Efficient Management in Smart Buildings », in *2018 IEEE Int. Conference on Environment and Electrical Engineering*, 2018, p. 1-6.

[2] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel & B. Wiswedel, « KNIME - the Konstanz Information Miner : Version 2.0 and Beyond », *ACM SIGKDD Explorations Newsletter* **11** (2009), n° 1, p. 26-31.

[3] M. R. Berthold & K.-P. Huber, « Missing Values and Learning of Fuzzy Rules », *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **06** (1998), n° 2, p. 171-178.

[4] J. H. Friedman, « Greedy function approximation : A gradient boosting machine. », *The Annals of Statistics* **29** (2001), n° 5, p. 1189-1232.

[5] M.-P. Gleizes, J. Boes, B. Lartigue & F. Thiébolt, « neOCampus : A Demonstrator of Connected, Innovative, Intelligent and Sustainable Campus », in *Intelligent Interactive Multimedia Systems and Services 2017*, vol. 76, Springer, 2018, p. 482-491.

[6] D. A. Guastella, « Dynamic learning of the environment for eco-citizen behavior », Phd thesis, Université Toulouse III – Paul Sabatier ; Università degli studi di Catania, 2020.

[7] D. A. Guastella, V. Camps & M.-P. Gleizes, « Estimating Missing Environmental Information by Contextual Data Cooperation », in *PRIMA 2019 : Principles and Practice of Multi-Agent Systems*, Springer, 2019, p. 523-531.

[8] ———, « A Cooperative Multi-Agent System for Crowd Sensing Based Estimation in Smart Cities », *IEEE Access* **8** (2020), p. 183051-183070.

[9] D. A. Guastella & E. Pournaras, « Cooperative Multi-Agent Traffic Monitoring Can Reduce Camera Surveillance », *IEEE Access* **11** (2023), p. 142125-142145.

[10] D. A. Guastella, N. Verstaevel, C. Valenti, B. Arshad & J. Barthélemy, « Evaluating Correlations in IoT Sensors for Smart Buildings », in *13th Int. Conference on Agents and Artificial Intelligence (ICAART)*, SCITEPRESS, 2021, p. 224-231.

[11] J. D. Hamilton, *Time Series Analysis*, Princeton University Press, 1994.

[12] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel & L. Thiele, « Deriving high-resolution urban air pollution maps using mobile sensor nodes », *Pervasive and Mobile Computing* **16** (2015), p. 268-285.

[13] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed & H. Chiroma, « The role of big data in smart city », *International Journal of Information Management* **36** (2016), nº 5, p. 748-758.

[14] T. K. Ho, « Random decision forests », in *Proceedings of 3rd Int. Conf. on Document Analysis and Recognition*, vol. 1, 1995, p. 278-282.

[15] J. Junttila, « Structural breaks, ARIMA model and Finnish inflation forecasts », *International Journal of Forecasting* **17** (2001), nº 2, p. 203 - 230.

[16] K. Kumar, M. Parida & V. K. Katiyar, « Short Term Traffic Flow Prediction for a Non Urban Highway Using Artificial Neural Network », *Procedia – Social and Behavioral Sciences* **104** (2013), p. 755-764.

[17] Y. Ma, S. Liu, G. Xue & D. Gong, « Soft Sensor with Deep Learning for Functional Region Detection in Urban Environments », *Sensors* **20** (2020), nº 12, p. 3348.

[18] T. Marwala, *Computational Intelligence for Missing Data Imputation, Estimation, and Management - Knowledge Optimization Techniques*, IGI Global, 2009.

[19] F. Mateo, J. J. Carrasco, A. Sellami, M. Millán-Giraldo, M. Domínguez & E. Soria-Olivas, « Machine learning methods to forecast temperature in buildings », *Expert Systems with Applications* **40** (2013), nº 4, p. 1061 - 1068.

[20] S.-V. Oprea & A. Bâra, « Machine Learning Algorithms for Short-Term Load Forecast in Residential Buildings Using Smart Meters, Sensors and Big Data Solutions », *IEEE Access* **7** (2019), p. 177874-177889.

[21] I. Pisa, I. Santín, J. Vicario, A. Morell & R. Vilanova, « ANN-Based Soft Sensor to Predict Effluent Violations in Wastewater Treatment Plants », *Sensors* **19** (2019), nº 6, p. 1280.

[22] V. Seal, A. Raha, S. Maity, S. Kumar Mitra, A. Mukherjee & M. K. Naskar, « A Simple Flood Forecasting Scheme Using Wireless Sensor Networks », *International Journal of Ad hoc, Sensor & Ubiquitous Computing* **3** (2012), nº 1.

[23] Z. Shan, Y. Xia, P. Hou & J. He, « Fusing Incomplete Multisensor Heterogeneous Data to Estimate Urban Traffic », *IEEE MultiMedia* **23** (2016), nº 3, p. 56-63.

[24] B. Spencer, O. Alfandi & F. Al-Obeidat, « A Refinement of Lasso Regression Applied to Temperature Forecasting », *Procedia Computer Science* **130** (2018), p. 728-735.

[25] D. Tomaras, V. Kalogeraki, N. Zvgouras, N. Panagiotou & D. Gunopulos, « Evaluating the Health State of Urban Areas Using Multi-source Heterogeneous Data », in *2018 IEEE 19th Int. Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, 2018, p. 14-22.

[26] B. Walczak & D. L. Massart, « The Radial Basis Functions – Partial Least Squares approach as a flexible non-linear regression technique », *Analytica Chimica Acta* **331** (1996), nº 3, p. 177-185.

[27] Y. Wang & I. H. Witten, « Modeling for optimal probability prediction », in *Proceedings of the 9th Int. Conference on Machine Learning*, Morgan Kaufmann, 2002, p. 650-657.

[28] I. H. Witten, F. Eibe & M. A. Hall, *Data Mining : Practical Machine Learning Tools and Techniques*, Elsevier, 2011.

[29] W. B. Wu, M. Woodroofe & G. Mentz, « Isotonic regression : Another look at the changepoint problem », *Biometrika* **88** (2001), nº 3, p. 793-804.

[30] L. Yu, N. Wang & X. Meng, « Real-time forest fire detection with wireless sensor networks », in *Proceedings of Int. Conf. on Wireless Communications, Networking and Mobile Computing, 2005*, vol. 2, 2005, p. 1214-1217.

[31] Y. Zhang & A. Haghani, « A gradient boosting method to improve travel time prediction », *Transportation Research Part C : Emerging Technologies* **58** (2015), p. 308 - 324.

[32] J. Y. Zhu, C. Sun & V. O. K. Li, « Granger-Causality-based air quality estimation with spatio-temporal (S-T) heterogeneous big data », in *Computer Communications Workshops (INFOCOM WKSHPS)*, IEEE, 2015, p. 612-617.

ABSTRACT. — The smart city aims at improving the quality of life of its citizens. Many sensors have to be deployed to monitor the state of the environment in which human activities take place. Despite these sensors being often cheap, their installation and maintenance costs increase rapidly with their number. In this paper, we address the problem of estimating environmental information where physical sensors are not available, to limit the costs related to the installation of additional sensors.

This paper presents the HybridIoT system for estimating missing environmental values in large-scale sensor networks. Our contribution is threefold: the definition of an approach for estimating missing values in large-scale environments, the definition and evaluation of a new geospatial method for estimating environmental values in the city of Toulouse, and finally the initial progress on deploying the system as part of the GIS neOCampus.

KEYWORDS. — Smart City, Cooperative Multi-Agent Systems, Missing Data Estimation.

*Manuscrit reçu le 16 mai 2023, révisé le 8 décembre 2023, accepté le 22 mars 2024.*