GreEn-ER Data - Jeux de données de consommation d'électricité dans le secteur tertiaire

Gustavo Felipe Martin Nascimento, Fréderic Wurtz, Benoit Delinchant, Patrick Kuo-Peng, Nelson Jhoe Batistela

GreEn-ER Data – Jeux de données de consommation d'électricité dans le secteur tertiaire

Gustavo Felipe Martin Nascimento[1,2*], Fréderic Wurtz[1], Benoit Delinchant[1], Patrick Kuo-Peng[2], Nelson Jhoe Batistela[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab, F-38000 Grenoble, France
[2] Université Fédérale de Santa Catarina, EEL, GRUCAD, Campus Universitaire João David Ferreira Lima, 88040-970, Florianópolis, Santa Catarina, Brésil
***gustavo-felipe.martin-nascimento@g2elab.grenoble-inp.fr**

RESUME. Les bâtiments jouent un rôle central dans la transition énergétique. La consommation d'électricité dans les bâtiments était de 63 % en 2017. Des études montrent qu'il est possible de réduire plus fortement la consommation dans les bâtiments quand on utilise des feedbacks directs (monitoring en temps réel) que par des feedbacks indirects (factures d'énergie). Ces feedbacks permettent de mieux comprendre comment le bâtiment consomme de l'énergie et donc de mieux agir. Dans cette perspective des méthodes de désagrégation d'énergie sont des outils puissants. Ces méthodes utilisent souvent des techniques d'apprentissage supervisé qui ont besoin de jeux de données conséquents. Ce travail contribue à la mise à disposition en open-access d'un jeu de données de consommation d'électricité d'un bâtiment tertiaire. Des notebooks avec métadonnées exploitent les mesures et analysent leur qualité tels que les données absentes ou aberrantes.

MOTS-CLÉS : jeux de données, bâtiment tertiaire, consommation d'électricité, qualité des données.

ABSTRACT. Buildings play a central role in the energy transition. Electricity consumption in buildings was 63% in 2017. Studies show that it is possible to reduce consumption in buildings more by using direct feedback (real-time monitoring) than by indirect feedback (energy bills) These feedbacks enable to understand better how the building consumes energy and thus to act better. From this perspective, energy disaggregation methods are powerful tools. These methods often use supervised machine learning techniques that require learning data sets. As datasets for the tertiary sector are rare, this work aims to contribute to the open-access availability of an electricity consumption dataset of a tertiary building. Notebooks with metadata evaluate the data and analyse its quality with regard to data gaps and outliers.

KEYWORDS : datasets,tertiary buildings, electricity consumption, data quality.

1. INTRODUCTION

L'énergie consommée dans les bâtiments correspond à une part importante de la consommation énergétique mondiale. En France, selon Bilan RTE 2018 (Réseau De Transport d'Électricité, 2019), environ 67,8 % de l'électricité est consommée dans les bâtiments, tant résidentiels que tertiaires. Aux États-Unis, la consommation d'électricité dans les bâtiments correspond à 74,7 %, selon les données de l'EIA (U.S. Energy Information Administration, 2019). Ces chiffres indiquent que les bâtiments jouent un rôle central dans la transition énergétique.

Des études montrent (Wood et Newborough, 2003) qu'il est possible de réduire plus efficacement la consommation d'énergie quand on dispose des informations en temps réel, en comparaison à de simple factures mensuelles. La mesure et le traitement des données de consommation ont donc une grande importance. Ces informations en temps réel peuvent être obtenus par une surveillance exhaustive, quand la plupart des charges sont mesurées individuellement, ou par des méthodes de *machine learning* qui utilisent des techniques de désagrégation de l'énergie. Ce dernier type de surveillance est connu sous le nom de NILM (Non-intrusive Load Monitoring) (Hart, 1992). Bien que la méthode NILM soit déjà apparue au siècle dernier, sa popularisation est plus récente due aux progrès informatiques dans les domaines de l'apprentissage automatique et de l'intelligence artificielle.

La plupart des recherches liées à la méthode NILM est basée sur des techniques d'apprentissage supervisé, dont les algorithmes nécessitent des données bien identifiées (Zoha et al, 2012). Il est donc important de disposer de données réelles de mesures de consommation d'électricité, afin que la recherche dans ce domaine puisse avancer. Il existe de jeux de données disponibles, avec par exemple des mesures globales et de sous compteurs, comme celles présentées par Kolter (Kolter et Johnson, 2011), Anderson (Anderson et al, 2012) et Makonin (Makonin et al 2013). Cependant, la plupart des jeux de données est dédié au secteur résidentiel et ceux dédiés au secteur tertiaire sont incomplets. Les équipements et profils de consommation dans ces deux secteurs sont très différents, il est donc important de disposer de jeux de données dédiés à des bâtiments du secteur tertiaire. C'est ce que nous proposons ici afin d'offrir à la communauté des chercheurs qui travaillent à l'analyse de la consommation d'électricité des bâtiments, des données de consommation exploitables accompagnées de leurs métadonnées.

2. QUALITE DES DONNEES

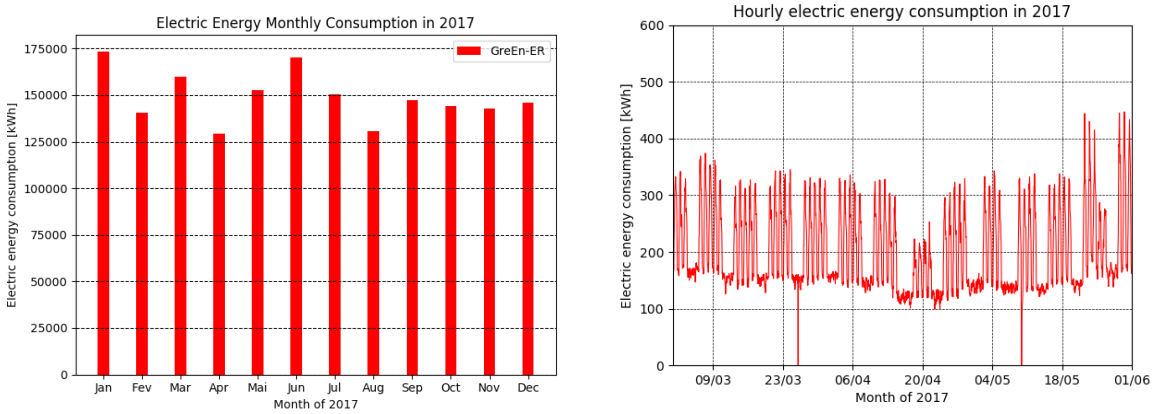
Pour réussir la mission d'utiliser les données de consommation en vue d'obtenir les informations en temps réel à partir des méthodes de désagrégation, il est primordial que les données répondent à des critères de qualité, faute de quoi les processus de *machine learning* avec séries temporelles contenant des anomalies peuvent aboutir à de mauvais modèles, car les paramètres et la variance du modèle sont affectés (Chang et al, 1988). Dans ce contexte, il est important de faire un prétraitement des données pour les nettoyer, améliorer leur qualité, et ainsi augmenter la précision des modèles qui peuvent en être issus.

Il y a plusieurs problèmes qui peuvent affecter les données de mesures et entraver leur analyse. En vue de signaler et classer ces problèmes, des caractéristiques particulières des données ont été définies et appelées "dimensions de la qualité des données". Les dimensions définis pour chaque jeu des données sont différentes en fonction des types de données utilisées. Dans le cas des séries temporelles, comme des données de consommation d'énergie, la complétude, l'actualité et l'exactitude sont parmi les principales dimensions. La complétude permet de mesurer si des données sont manquantes, l'exactitude mesure si les échantillons sont corrects et fiables et l'actualité mesure si les informations sont à jour, tandis que les *outliers* sont des données qui semblent incompatibles avec le reste de la série, par exemple par un éloignement statistique. En fonction du type des données présentées dans ce travail, nous abordons la détection de données aberrantes (*outliers*), l'analyse de complétude et d'exactitude.

3. LE BATIMENT GREEN-ER

Le bâtiment GreEn-ER est situé au Polygone Scientifique, à la Presqu'ile de Grenoble. Il regroupe l'école d'ingénierie Grenoble-INP Ense3, le laboratoire G2Elab et, aussi, des plateformes de formation

et recherche. Le bâtiment a plus de 22.000 m2 de surface qui sont divisés sur 6 étages et la toiture. Il y a environ 1.500 étudiants et quelques centaines de professeurs, chercheurs et personnels qui le fréquentent. Comme il s'agit d'un grand bâtiment, sa consommation d'électricité est aussi importante. Dans des journées typiques, la puissance active peut s'élever à plus de 300 kW. C'est un bâtiment massivement surveillé et contrôlé avec plus de 1.500 capteurs dont environ 330 compteurs d'énergie électrique. Ces derniers mesurent la consommation des différentes charges du bâtiment, comme l'éclairage, les prises de courant, les Centrales de Traitement d'Air (CTAs), les groupes froids, les pompes, etc. (Delinchant et al., 2016). Les données mesurées sont utilisées pour contrôler les conditions internes et pour suivre la consommation. Les figures suivantes montrent la consommation mensuelle de l'année 2017 et la consommation horaire, entre les mois de février et juin, dont il est possible de voir le profil de consommation de l'énergie électrique. Il est important de rappeler ici que le bâtiment n'est pas chauffé à l'énergie électrique mais par le réseau de chaleur de Grenoble. Les autres usages sont électrique, c'est par exemple le cas du rafraichissement en été assuré par des pompes à chaleur. Malgré tout, une faible dépendance saisonnière peut être constatée sur la première figure. Par contre, une variation quotidienne et hebdomadaire est clairement visible sur la deuxième figure.



a) Consommation mensuelle d'électricité en 2017

b) Courbe de charge entre les mois de février et juin 2017

Figure 1 : Consommation d'électricité du bâtiment GreEn-ER en 2017.

Le bâtiment reçoit l'énergie électrique sous une tension de 20 kV et, cette tension est abaissée à 410 V à l´aide de deux transformateurs. Chaque transformateur est lié à un TGBT (Tableau Générale de Basse Tension) dédié, appelés respectivement TGBT1 et TGBT2. Ces deux tableaux alimentent les tableaux de distribution qui, à leur tour, distribuent l'électricité aux charges du bâtiment. La consommation de l'électricité des tableaux et des charges est mesurée par des compteurs dédiés dont les mesures constituent les jeux des données de consommation présentés dans ce travail.

4. LES JEUX DE DONNEES GREEN-ER

Afin de constituer les jeux des données de consommation de l'énergie électrique, tous les compteurs d'électricité pour les années 2017 et 2018 ont été recueillies. Les données mesurées sont disponibles dans des fichiers CSV (valeurs séparées par des virgules). Un fichier est mis à disposition pour chaque compteur d'électricité, pour chaque année d'analyse, soit environ 650 fichiers distincts au total. Les métadonnées, qui constituent un ensemble d'information permettant de comprendre la signification des

données, sont disponibles dans les fichiers Jupyter Notebook (ipynb). Un notebook est un document qui permet de combiner du texte et d'image avec des codes de programmation. Nous travaillons avec des notebook Jupyter en langage Python. Ainsi, il est possible de présenter le bâtiment, d'illustrer la division du système électrique et de faire une brève exploration des données. En raison de la complexité du système et de la quantité de données disponibles, quatre fichiers Jupyter Notebook différents ont été préparés. L'un d'eux explore la consommation totale du bâtiment, un autre explore le TGBT1 et un autre explore le TGBT2. Enfin, un autre notebook explore les données de la plateforme PREDIS-MHI, une partie isolée énergétiquement du reste du bâtiment. Le tableau suivant présente la quantité de fichiers que chaque Notebook exploite.

Notebook	Nombre de fichiers
GreEn-ER Consommation Générale	4
TGBT1	113
TGBT2	120
PREDIS-MHI	91

Tableau 1 : Nombre des fichiers exploités par chaque notebook.

Chaque notebook décrit la partie qu'il explore, avec des plans du bâtiment, des schémas pour illustrer son système électrique et des tableaux qui définissent à quelle charge chaque compteur est lié. Grâce au choix de l'année et du compteur à explorer par l'utilisateur, le notebook calcule la puissance moyenne et de la consommation de l'année choisie. Il montre aussi des courbes de charge interactives et la consommation mensuelle sous forme graphique. Un exemple de ces brèves analyses a déjà été montré sur la Figure 1.

Les jeux de données sont mis à disposition sur la plateforme ouverte Mendeley Data (Martin Nascimento et al, 2020). Il est important de souligner que les données qui constituent ces jeux de données sont brutes, c´est à dire sans aucun prétraitement, de telle sorte que des problèmes de qualité des données sont présents, ouvrant la possibilité aux chercheurs d'utiliser les jeux de données également pour des recherches dans ce domaine. La section suivante vise à présenter des exemples de quantification de la qualité des données des jeux de données créés.

5. QUALITE DES DONNEES GREEN-ER

En vue de quantifier la qualité du jeu des données créé pour le bâtiment GreEn-ER par rapport aux dimensions choisies (quantité des *outliers*, complétude et exactitude), il est nécessaire de définir des règles plus adaptées aux données utilisées vu que des données mesurées pour chaque type de capteur peut présenter des caractéristiques différentes. Ces règles sont détaillées dans les sections suivantes. Un script python est aussi disponible et joint aux jupyter notebooks avec l'objectif d'exploiter la qualité des données.

5.1. QUANTITE DES *OUTLIERS*,

Les *outliers* sont des données qui semblent incompatibles avec le restant de la série et sont statistiquement éloignés des autres échantillons. On peut observer un exemple sur la Figure 2.

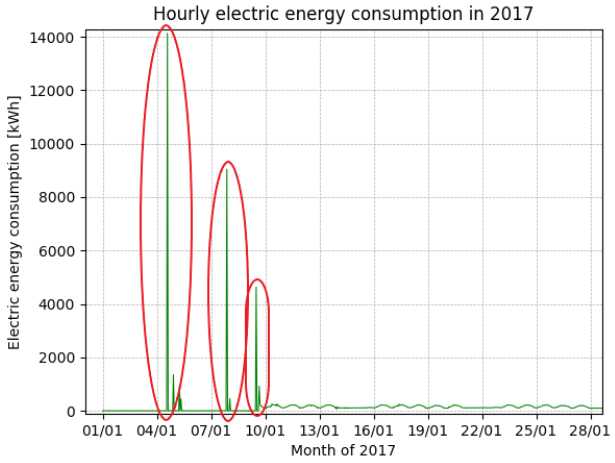


Figure 2 : Exemple de la présence d'outliers.

Il existe plusieurs méthodes pour identifier ce type de problème. Dans ce travail, l´utilisation d´une méthode similaire à celle décrite par Tukey (Tukey, 1977) a été choisie. Dans celui-ci, les valeurs des séries sont classées par ordre croissant. Puis les valeurs du premier (25%) et du troisième quartile (75%) sont calculées. Avec ces valeurs en main, l'intervalle interquartile est calculé à l'aide de l'équation 1, dans laquelle IQR représente cet intervalle et Q_1 et Q_3 représentent la valeur du premier et du troisième quartile, respectivement.

$$IQR\ Q\square\ \square_3\ Q_1 \quad \textbf{(1)}$$

Après le calcul de l'intervalle interquartile, la limite supérieure est calculée sur la base de l'équation 2, dans laquelle cette limite est représentée par UPB. La limite inférieure est égale à zéro, car les valeurs de consommation négatives ne sont pas possibles.

$$UPB\ Q\square\ \square_3\ (3IQR) \quad \textbf{(2)}$$

Dans plusieurs séries appartenant au jeu de données créé, la charge surveillée reste inactive pendant longtemps, de sorte que les valeurs nulles sont souvent supérieures à 75%. Pour cette raison la méthode décrite ci-dessus a été adaptée en modifiant les quartiles utilisés à 15 % et 85 %. Quelques exemples de résultats sont présentés sur le tableau suivant.

Compteur	Limite Supérieure	Nombre d'*outliers*
Td CEE-BT-789-ARM-BAT-F19-ARM_CTA_EST-TGBT_1	68	60
Td SC-BT-779-ARM-BAT-F09-TD1_EE3_R1-TGBT_1	72	63
Td SC-BT-771-ARM-BAT-TGBT_1	462	6
Td SC-BT-823-FRD-BAT-GF1-TD_GF	96	118

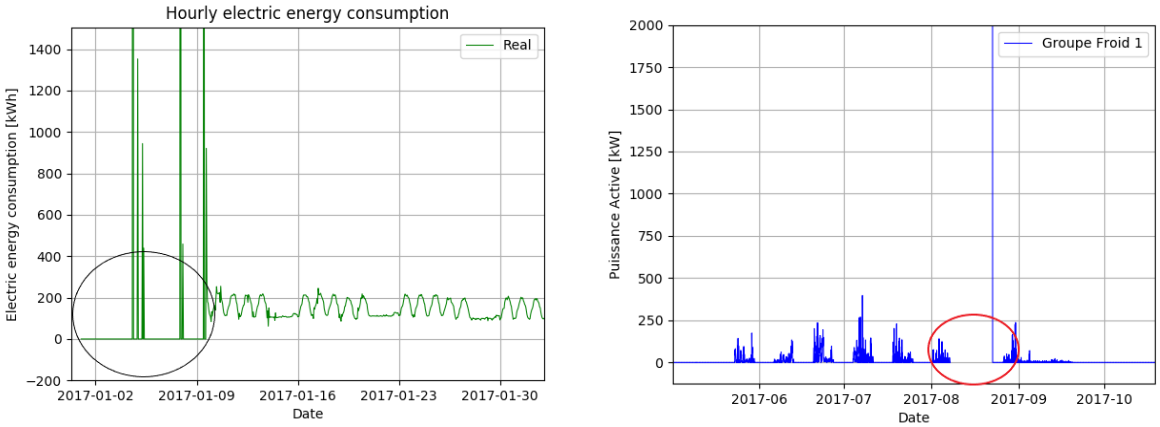
Tableau 2 : Exemples de résultats liés au nombre d'outliers des données en 2017.

5.2. COMPLETUDE

Les données qui constituent les jeux de données créés peuvent présenter deux types de problèmes de complétude. Le premier, arrive quand l'échantillon ne présent aucune valeur. Donc, pour quantifier le nombre d'échantillons liés à ce type de problème il suffit de trouver les trous dans la série temporelle. Le deuxième type s'agit d'une caractéristique particulière des données de puissance recalculées à partir des mesures de compteurs d'énergie. Quand il y a un problème de communication entre les compteurs et le système d'archivage, le système de gestion technique du bâtiment (GTB) sauvegarde ces données comme une valeur nul. Après le rétablissement de la communication, le système GTB sauvegarde la

valeur de la consommation totale de la période sans communication en un unique instant. Cette valeur devient un *outlier* car très différentes des autres valeurs habituelles. Alors, pour bien quantifier ce type de problème il faut détecter le nombre de zéros consécutifs suivis d'un *outlier*. Ce type de problème peut être visualisé sur la Figure 3.

Le Tableau 3 présente des exemples de résultats de complétude pour quelques compteurs. Les résultats pour les autres compteurs peuvent être calculés avec le script python disponible joint aux jeux de données. Les chiffres présentés sur ce tableau montrent qu'il y a des problèmes de complétude des données. Cependant, on peut percevoir que la plupart des échantillons classés comme manquants sont liés à des problèmes de communication entre le compteur et le système d'archivage.



a) Exemple des zéros liés au manque de communication dans la série temporelle
b) Exemple de trou dans la série temporelle.

Figure 3 : Exemple de problèmes de qualité de données liés à la complétude.

Compteur	Échantillons inexistants	Zéros suspects	Échantillons suspects inexistants	Total d'échantillons	Complétude [%]
Td CEE-BT-789-ARM-BAT-F19ARM_CTA_EST-TGBT_1	1	620	621	8785	92,93%
Td SC-BT-779-ARM-BAT-F09TD1_EE3_R1-TGBT_1	1	1111	1112	8785	87,34%
Td SC-BT-771-ARM-BAT-TGBT_1	1	197	198	8785	97,75%
Td SC-BT-823-FRD-BAT-GF1-TD_GF	362	409	771	8785	91,22%

Tableau 3 : Exemples de résultats liés à la complétude des données en 2017.

5.3. EXACTITUDE

En vue d'évaluer l'exactitude des données mesurées, une comparaison entre les mesures des compteurs qui sont liés de manière hiérarchique a été faite. Un compteur principal a été défini et les valeurs mesurées avec la somme des sous-compteurs liés à ce compteur principal ont été comparées. Un exemple de cette liaison est illustré sur la Figure 4. Dans cet exemple, la mesure du compteurs « 840 » avec la somme des sous-compteurs « 823 », « 824 » et « 825 » est comparée. Ces compteurs sont responsables pour mesurer la consommation d'un système de refroidissement du bâtiment.

Il a été établi qu'un échantillon est correcte quand la variation en pourcentage est plus petite que la tolérance admissible. Le nombre d'échantillons corrects pour des tolérances admissibles allant de 10% à 100% a donc été calculé. Les résultats obtenus sont présentés sur le Tableau 4.

Sur la base des résultats présentés ci-dessus, on peut constater que l'exactitude des échantillons, telle que définie dans ce travail, suit une tendance presque linéaire en fonction de la tolérance admissible jusqu'au niveau de 90%. Pour d'autres exemples cette tendance présente un profil logarithmique, ce que l'on peut constater en utilisant le script python dédié à la qualité des données. Toutefois, les moyennes des grandeurs comparées dans la période évaluée ne diffèrent pas beaucoup. La puissance moyenne du compteur principal était de 13,3 kW, tandis que la moyenne de la somme des compteurs secondaires était de 13,7 kW. Cependant, l'écart type est tout à fait différent. Alors que l'écart-type de la série correspondant au compteur principal est de 21,6 kW, l'écart type de la somme des compteurs secondaires est de 94,7 kW.

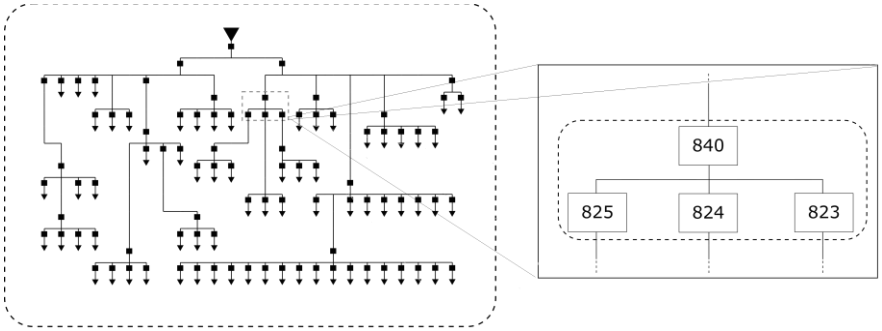


Figure 4 : Partie du système électrique du bâtiment pour calculer l'exactitude des données.

Tolérance [%]	Nombre d'échantillons corrects	Nombre d'échantillons	Exactitude [%]
10%	2877	8785	32,7%
20%	3566	8785	40,6%
30%	3929	8785	44,7%
40%	4441	8785	50,6%
50%	4854	8785	55,3%
60%	5190	8785	59,1%
70%	5473	8785	62,3%
80%	5791	8785	65,9%
90%	6002	8785	68,3%
100%	7362	8785	83,8%

Tableau 4 : Exemples de résultats liés à l'exactitude des données en 2017.

6. CONCLUSIONS ET PERSPECTIVES

Différent de la plupart des jeux de données existants sur la littérature, les jeux de données créés et présentés dans ce travail sont constitués des données complètes de consommation d'électricité d'un bâtiment tertiaire massivement surveillé. Les données sont disponibles sur une plateforme ouverte, en permettant l'accès à tous les chercheurs intéressés. Ces jeux de données peuvent être utilisés sur plusieurs voies de recherche tels que la désagrégation d'énergie, le prétraitement des données, la qualité des données, etc.

Pour une analyse préliminaire nous avons quantifié la qualité des données de quelques compteurs en regardant les dimensions de complétude, exactitude et la présence des *outliers*. Les résultats obtenus

montrent qu'il est nécessaire d'exécuter un prétraitement des données pour les utiliser comme jeux de données d'apprentissage pour des algorithmes de *machine learning*.

En perspective, et en complément du jeu de données brutes mis à disposition, il est intéressant de développer des algorithmes de prétraitement et de nettoyage de ces données, afin de produire un jeu directement exploitable pour de l'apprentissage par exemple.

7. BIBLIOGRAPHIE

Anderson, K., Ocneanu, A., Benitez, D., Carlson, D., Rowe, A. et Berges, M. 2012. « BLUED: A fully labelled public dataset for event-based non-intrusive load monitoring research". *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD), 1-5.*

Bansal, S.; Schmidt, M. 2017 « Energy Disaggregation Methods for Commercial Buildings Using Smart Meter and Operational Data ». *AAAI Workshops*.

Chang, I., Tiao, G. C., Chen, C., 1988. « Estimation of time series parameters in the presence of outliers. » *Journal of Technometrics 30 (2), 193–204.*

Delinchant, B., Wurtz, F., Ploix, S., Schanen J. et Marechal Y., 2016 « GreEn-ER living lab: A green building with energy aware occupants, » *2016 5th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)*, Rome, 2016, pp. 1-8.

Hart, G. W. 1992 « Nonintrusive appliance load monitoring » *Proc. IEEE, vol. 80, no. 12, pp. 1870–1891*

Kolter, J.Z.; Johnson, M.J. 2011 « REDD: A Public Data Set for Energy Disaggregation Research. » *In Proceedings of the SustKDD Workshop on Data Mining Applications in Sustainability*, San Diego, CA, USA, August 2011; pp. 1–6.

Makonin, S., Popowich, F., Bartram, L., Gill, B. & Bajic, I. V. 2013. « AMPds: A public dataset for load disaggregation and eco-feedback research ». *Electrical Power and Energy Conference (EPEC), 2013 IEEE, 1-6.*

Martin Nascimento, G. F., Delinchant, B., Wurtz, F., Kuo-Peng, P., Jhoe Batistela, N., et Laranjeira, T., "GreEn-ER - Electricity Consumption Data of a Tertiary Building", *Mendeley Data*, V1, 2020 http://dx.doi.org/10.17632/h8mmnthn5w.1

RÉSEAU DE TRANSPORT D'ÉLECTRICITÉ. 2019 Bilan Électrique 2018. RTE - Direction innovation et données. url :*https://www.rte-france.com/sites/default/files/be_pdf_2018v3.pdf*.

TUKEY, John Wilder. « Exploratory data analysis **».** *Reading: Addison-Wesley, c1977. 688 p. ISBN 0201076160*

U.S. ENERGY INFORMATION ADMINISTRATION. 2019 « Table 5.1. Sales of Electricity to Ultimate Customers. Electric Power Monthly. » *https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_5_01*.

Wood, G. and Newborough, M. (2003) « Dynamic Energy-Consumption Indicators for Domestic Appliances: Environment, Behavior and Design. » *Energy and Buildings, Volume 35, Issue 8, September 2003, Pages 821-841*

Zoha, Ahmed, Gluhak, Alexander, Imran, Muhammad et Rajasegarar, Sutharshan. (2012). « NonIntrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey.” *Sensors (Basel, Switzerland). 12. 16838-16866. 10.3390/s121216838.*