



**HAL**  
open science

# Optimization of resources and consumption of smart buildings for energy efficiency

Gustavo Felipe Martin Nascimento

► **To cite this version:**

Gustavo Felipe Martin Nascimento. Optimization of resources and consumption of smart buildings for energy efficiency. Electric power. Université Grenoble Alpes [2020-..]; Universidade federal de Santa Catarina (Brésil), 2022. English. NNT : 2022GRALT078 . tel-04008796

**HAL Id: tel-04008796**

**<https://theses.hal.science/tel-04008796v1>**

Submitted on 28 Feb 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## THÈSE

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES**

**préparée dans le cadre d'une cotutelle entre  
l'Université Grenoble Alpes et la Universidade  
Federal de Santa Catarina**

Spécialité : **Génie Électrique**

Arrêté ministériel : le 25 mai 2016

Présentée par

**Gustavo Felipe MARTIN NASCIMENTO**

Thèse dirigée par **Frédéric WURTZ** et **Patrick KUO-PENG**  
codirigée par **Benoit DELINCHANT** et **Nelson Jhoé BATISTELA**

préparée au sein des Laboratoires : **Laboratoire de Génie  
électrique de Grenoble (G2Elab)** et **Grupo de Concepção e  
Análise de Dispositivos Eletromagnéticos (GRUCAD)**

dans l'École Doctorale **Electronique, Electrotechnique,  
Automatique, Traitement du Signal (EEATS)** et **Programa de  
Pós-Graduação em Engenharia Elétrica – UFSC**

## **Optimisation des ressources et consommation des « smart buildings » en vue de l'efficacité énergétique**

Thèse soutenue publiquement le **18 octobre 2022**  
devant le jury composé de :

**Monsieur Stéphane PLOIX**

Professeur d'université, Université Grenoble Alpes, Président

**Madame Florence OSSART**

Professeure d'université, Sorbonne Université, Rapporteur

**Monsieur José Roberto CARDOSO**

Professeur d'université, USP, Rapporteur

**Monsieur Nelson SADOWSKI**

Professeur d'université, UFSC, Examineur

**Monsieur Frédéric WURTZ**

Directeur de Recherche, Grenoble-INP, Membre

**Monsieur Patrick KUO-PENG**

Professeur d'université, UFSC, Membre





## Acknowledgements

Firstly, I would like to thank all member of the jury committee that read carefully this work, proposed improvements and made a very productive discussion during the defense: Mr. Stéphane PLOIX, Mrs. Florence OSSART, Mr. José Roberto CARDOSO and Mr. Nelson SADOWSKI.

This thesis is the result of a collaboration between two renowned universities: the University Grenoble Alpes, within the G2Elab, Grenoble-INP, in Grenoble, France and the Federal University of Santa Catarina within the and the GRUCAD, in Florianópolis, Brazil. Therefore, I would also like to thank all the staff from both universities that made possible this collaboration. Also, I would like to deeply thank the Carnot Énergies du Futur Institute, that funded this research.

This thesis would not be accomplished without the supervision and guidance from my two thesis directors, Frédéric WURTZ and Patrick KUO-PENG, who were always there whether for ideas, directions to follow or even for moral support even during the toughest moments. At the same time, I would also like to thank Benôit DELINCHANT e Nelson JOE BATISTELA, my two co-directors for all the support during this period.

When we live abroad, our friends become our family. So, I would also like to thank the friends that I made during the time spent in Grenoble and that I will always deeply cherish. Pedro VON HOHENDORFF, Sabine FRELLO, Lucas DE SOUZA, Alexa PORTELA, Arthur FEUERHARMEL, Beatriz DE LUCA, Bruna ANDREOTTI, Vinicius SOUZA, Vitória TORRES, Evan THOMPSON, Caio FREITAS, Maria Luiza PEDROSA and Andre ANDRETTA.

However, it was really difficult for me to leave my family in Brazil to pursue the dream of becoming a doctor. Therefore, I would also like to thank all my family, especially my parents, Rubens and Maria Elisete, my siblings, Fernando, Vinicius and Laura, and my grandmothers, Maria Ivone and Gisela, that sadly passed away during the final stages of this PhD and couldn't see me finish it.

Finally, and most important, I would like to thank my lovely wife, Priscila, who agreed to face this adventure with me. You know that this degree is not only mine, but ours. Without you nothing of this would be possible, and surely I would be home sooner, without the degree, or facing the Isère. I love you.

Thank also to you that is interested to read this work.

## Table of Contents

<b>1</b>	<b>Scientific Context – Electricity Consumption in Buildings</b>	<b>21</b>
1.1	Overview of the Worldwide Energy Consumption and GHG emissions	22
1.2	Electricity Consumption in Buildings	27
1.3	Conclusions	34
<b>2</b>	<b>Energy Consumption Datasets of Buildings – Objectives, State of Art and The Green-ER Case</b>	<b>37</b>
2.1	The Importance of Datasets Towards the Research Advance	38
2.2	Data Valorization – Towards Open-Science	39
2.3	The GreEn-ER Dataset	42
2.4	GreEn-ER Electricity Consumption	49
2.5	Conclusions	63
<b>3</b>	<b>Data Quality in Energy Domain – The GreEn-ER Case</b>	<b>67</b>
3.1	Overview of Data Quality Problems	68
3.2	Data Quality Assessment of the GreEn-ER Dataset	69
3.3	Poor Data Quality Influence in Machine Learning	76
3.4	Conclusions	79
<b>4</b>	<b>Outlier Detection in Buildings’ Power Consumption Data Using Forecast Error</b>	<b>83</b>
4.1	Forecast Error Method Overview	84
4.2	Statistical Methods for Outlier Detection	85
4.3	Results and Analysis	90
4.4	Assessing Completeness of GreEn-ER loads	99
4.5	Conclusions	101
<b>5</b>	<b>Non-Intrusive Load Monitoring – State of Art</b>	<b>105</b>
5.1	Direct and Indirect Feedbacks Towards Energy Consumption Savings	106
5.2	NILM Methods Based On the Activation of Finite States	109
5.3	NILM Methods Based on Artificial Neural Networks	112
5.4	The NILM Toolkit – NILMTK	121
5.5	Conclusions	124
<b>6</b>	<b>Disaggregation methods applied to a tertiary building environment – The GreEn-ER case</b>	<b>129</b>
6.1	Energy Disaggregation in the Tertiary Sector	130
6.2	Energy Disaggregation Applied to GreEn-ER loads	132
6.3	Conclusions	143
<b>7</b>	<b>Energy Audits Using Energy Disaggregation – The GreEn-ER Compressed Air System Case</b>	<b>147</b>
7.1	Disaggregation and Energy Audits Overview	148
7.2	Energy Audits Overview	149
7.3	Case Studies – Estimating Compressed Air Leakage	152

---

7.4	Conclusions	180
8	General Conclusions and Directions for Future Works	185
8.1	General Conclusions	186
8.2	Major Contributions	189
8.3	Future Directions	190
8.4	Publications	191



## **Abstract**

In the challenges of energy transition and reducing consumption, buildings play a key role, as they consume a significant part of worldwide energy consumption. Considering electrical energy alone about 50% is consumed in residential or tertiary buildings. This figure indicates the great potential for energy savings in buildings. The residential sector accounts for 37% of consumption, while electricity consumption in tertiary buildings accounts for 32% of the total. This superior consumption in residential buildings has led to an explosion of studies on the energy consumption of these buildings, so that the tertiary sector still has much to be explored.

The first step in reducing energy consumption is to conduct an energy audit, which is a detailed evaluation of the energy performance of the systems in a facility. It makes easier identifying and quantifying the energy savings potentials. Thus, this thesis studies methods to enhance energy audits in buildings, especially those from the tertiary sector. These studies involve the comprehensive analysis of data from the GreEn-Er building at Grenoble-INP, a smart-building exhaustively monitored offering an extremely rich and promising field for experimentation and extrapolation.

The research in energy consumption of tertiary buildings presents yet a deficit when compared to the residential sector. This can be associated to the lack of public available datasets of commercial buildings, when compared to the residential sector. Hence, to address this issue when decrease this gap, this thesis presents the GreEn-ER dataset, containing aggregated and disaggregated electricity consumption data from the building.

Nevertheless, real data bring real problems and the data extracted from the GreEn-ER dataset is no exception. Hence, this thesis also addresses data quality metrics, especially in terms of accuracy and completeness. Additionally, a method for identifying outliers was developed during the course of the thesis.

Furthermore, in the context of energy audits, NILM method can be useful for enhancing the analysis performed. Hence, the present works also addresses the application of NILM algorithms to loads typically present in tertiary buildings, since most of the techniques developed to date have been developed regarding residential buildings. Finally, an enhancement to an energy audit, by quantifying the compressed air leakage using an energy disaggregation method is presented.





## Résumé

Depuis qu'ils consomment une part importante de la consommation énergétique mondiale, les bâtiments jouent un rôle central dans le défis de la transition énergétique et de la réduction de la consommation. Si l'on considère l'énergie électrique seule, environ 50% est consommée dans les bâtiments résidentiels ou tertiaires. Ce chiffre indique l'important potentiel d'économies d'énergie dans les bâtiments. Le secteur résidentiel représente 37% de la consommation, tandis que la consommation d'électricité dans les bâtiments tertiaires représente 32% du total. Cette consommation supérieure dans les bâtiments résidentiels a entraîné une explosion des études sur la consommation énergétique de ces bâtiments, de sorte que le secteur tertiaire a encore beaucoup à explorer.

Un premier pas pour réduire la consommation d'énergie c'est la réalisation d'une audit énergétique, qui est une évaluation détaillée de la performance énergétique des systèmes d'une installation. Il permet d'identifier et de quantifier plus facilement les potentiels d'économies d'énergie. Ainsi, cette thèse étudie les méthodes permettant d'améliorer les audits énergétiques dans les bâtiments, notamment ceux du secteur tertiaire. Ces études impliquent l'analyse complète des données du bâtiment GreEn-Er à Grenoble-INP, un smart-building massivement surveillé offrant un champ d'expérimentation et d'extrapolation extrêmement riche et prometteur.

La recherche sur la consommation énergétique des bâtiments tertiaires présente encore un déficit par rapport au secteur résidentiel. Cela peut être associé au manque d'ensembles de données publiques disponibles sur les bâtiments commerciaux, par rapport au secteur résidentiel. C'est pourquoi, afin d'aborder ce problème et de combler cette lacune, cette thèse présente le jeu de données GreEn-ER, qui contient des données agrégées et désagrégées sur la consommation d'électricité des bâtiments.

Cependant, les données réelles emportent des problèmes réels et les données extraites du jeu de données GreEn-ER ne font pas exception. Par conséquent, cette thèse aborde également les métriques de qualité des données, notamment en termes de précision et de complétude. De plus, une méthode d'identification des outliers a été développée au cours de cette thèse.

En outre, dans le contexte des audits énergétiques, la méthode NILM peut être utile pour améliorer l'analyse effectuée. Ainsi, les présents travaux abordent également l'application des algorithmes NILM aux charges typiquement présentes dans les bâtiments tertiaires, puisque la plupart des techniques développées à ce jour l'ont été pour les bâtiments résidentiels. Enfin, une amélioration d'un audit énergétique, en quantifiant les fuites d'air comprimé à l'aide d'une méthode de désagrégation énergétique est présentée.



## List of Figures

- Figure 1 – Evolution of the worldwide life expectancy at birth. [1]
- Figure 2 – Global GHG emissions by sector. [2]
- Figure 3 – Annual primary energy consumption in 2019 in selected countries of the EU.
- Figure 4 – Distribution of worldwide electricity consumption by sector.
- Figure 5 – Distribution of electricity consumption by sector in France.
- Figure 6 – Distribution of consumption by use in the French residential sector in 2019.
- Figure 7 – Distribution of consumption by activity of tertiary buildings in France in 2019.
- Figure 8 – Distribution of consumption by appliance type of tertiary buildings in France in 2019.
- Figure 9 – The GreEn-ER building [22].
- Figure 10 – Map indicating the location of the building (21 Avenue des Martyrs, 38000, Grenoble, France).
- Figure 11 – Electric scheme of the GreEn-ER building.
- Figure 12 – Electric scheme of the GreEn-ER building as presented by the BMS.
- Figure 13 – On-line client of the BMS showing the parameters of the room 4-A-017, located at the fourth floor.
- Figure 14 – Structure of the main folder of the dataset.
- Figure 15 – Structure of a folder of a specific year of the dataset.
- Figure 16 – Overview of the PREDIS-MHI [22].
- Figure 17 – Electric scheme of the TGBT1 present in it respective notebook.
- Figure 18 – Electric scheme of the switchboards connected to TGBT1 present in it respective notebook.
- Figure 19 – Example of the floor plan of the 2<sup>nd</sup> floor of the GreEn-ER building with zones and their respective switchboards.
- Figure 20 – Monthly electricity consumption of the GreEn-ER building in 2017.
- Figure 21 – Load curve of the GreEn-ER building in 2017.
- Figure 22 – Zoom in on the load curve of the GreEn-ER building in 2017.
- Figure 23 – Electricity distribution between TGBT1 and TGBT2.
- Figure 24 – Electricity distribution among main loads of TGBT1.
- Figure 25 – Zoom in on the load curve of the Datacenter and heat pumps system.
- Figure 26 – Zoom in on the load curve of the Ense<sup>3</sup> and G2Elab areas.
- Figure 27 – Zoom in on the load curve of the AHUs.
- Figure 28 – Electricity distribution among main loads of TGBT2.
- Figure 29 – Zoom in on the load curve of the TD-GF and Crous.
- Figure 30 – Zoom in on the load curve of the Predis and common areas.
- Figure 31 – Zoom in on the load curve of the Predis-MHI and other loads.
- Figure 32 – Zoom in on the load curve of the air compressors.
- Figure 33 – Annual electricity consumption of the air compressors from 2017 to 2020.
- Figure 34 – Monthly electricity consumption of the air compressors from 2017 to 2020.
- Figure 35 – Average electricity consumption of each weekday of the air compressors from 2017 to 2020.
- Figure 36 – Compressed air system load curve during vacation time in 2018.
- Figure 37 – Zoom in of the TGBT1 load curve highlighting some data quality problems.
- Figure 38 – Real and virtual meters of TGBT2 in terms of power.
- Figure 39 – TGBT2's virtual meter accuracy as a function of the sampling interval and the tolerance allowed
- Figure 40 – Location of the Groupe Froid 1 meter.

- Figure 41 – Load curve of the Groupe Froid 1 highlighting a gap in the data.
- Figure 42 – Schema of the decision tree algorithm.
- Figure 43 – Schema of the random forest algorithm. [15]
- Figure 44 – Global load curve of GreEn-ER in 2017 used as training data.
- Figure 45 – Forecast results for 2018 using data with DQ problems as training data set.
- Figure 46 – Forecast results for 2018 using healthy data as training set.
- Figure 47 – Percentages in normal distribution between standard deviations. Based on Straker [19].
- Figure 48 – Example of a box-and-whisker plot for a normal distribution. Based on Olano et al. [22].
- Figure 49 – Synthetic GreEn-ER global power consumption.
- Figure 50 – Synthetic GreEn-ER global power consumption with inserted outliers.
- Figure 51 – Regression results using the random forest algorithm on adapted data.
- Figure 52 – Data features’ importance in the random forest regression for the adapted data series.
- Figure 53 – Real GreEn-ER global power consumption with inserted outliers.
- Figure 54 – Regression results using the random forest algorithm on real data
- Figure 55 – Data features importance in the random forest regression of the real data series.
- Figure 56 – TGBT1 load curve.
- Figure 57 – Energy saving potential of different types of feedback, figure reproduced from [1].
- Figure 58 – Illustration of the Hidden Markov Model. Based on Bonfigli and Squartini. [21]
- Figure 59 – Illustration of the Factorial Hidden Markov Model. Based on Bonfigli and Squartini. [21]
- Figure 60 – Representation of an Artificial Neuron. Based on [14]
- Figure 61 – Representation of a Fully Connected Feed-Forward Artificial Neural Network with three hidden layers.
- Figure 62 – Representation of a Convolutional Artificial Neural Layer for image processing [23].
- Figure 63 – The conceptual structure of a convolutional neural network used for the purpose of time-series analysis. [27]
- Figure 64 – Representation of a Recurrent Artificial Neural Layer. [26]
- Figure 65 – Representation of a LSTM cell. [24].
- Figure 66 – Representation of a Gate Recurrent Unit cell. [25]
- Figure 67 – Artificial Neural Network based on LSTM layers for energy disaggregation. [10]
- Figure 68 – Artificial Neural Network based on GRU layers for energy disaggregation. [19]
- Figure 69 – Sequence-to-point Artificial Neural Network for energy disaggregation. [20]
- Figure 70 – YAML files containing GreEn-ER dataset metadata for the integration to the NILMTK.
- Figure 71 – Electric schema of the TGBT2 with the target loads highlighted.
- Figure 72 – Load curve of the Air Compressors switchboard and the CROUS switchboard.
- Figure 73 – Load curve of the TD-GF switchboard and the CROUS switchboard.
- Figure 74 – Load curve of the aggregated load “Fantome”.
- Figure 75 – Disaggregation results of the air compressor using combinatorial optimization.
- Figure 76 – Disaggregation results of the CROUS using combinatorial optimization.
- Figure 77 – Disaggregation results of the TD-GF using combinatorial optimization.
- Figure 78 – Disaggregation results of the “fantôme” load using combinatorial optimization.
- Figure 79 – Disaggregation results of the Air Compressor using FHMM.
- Figure 80 – Disaggregation results of the CROUS using FHMM.
- Figure 81 – Disaggregation results of the TD-GF using FHMM.
- Figure 82 – Disaggregation results of the “Fantôme” load using FHMM.
- Figure 83 – Disaggregation results of the Air Compressor using LSTM architecture.
- Figure 84 – Disaggregation results of the CROUS using LSTM architecture.
- Figure 85 – Disaggregation results of the TD-GF using LSTM architecture.
- Figure 86 – Disaggregation results of the “Fantôme” load using LSTM architecture.
- Figure 87 – Disaggregation results of the Air Compressor using GRU architecture.
- Figure 88 – Disaggregation results of the CROUS using GRU architecture.

- Figure 89 – Disaggregation results of the TD-GF using GRU architecture.
- Figure 90 – Disaggregation results of the “Fantome” load using GRU architecture.
- Figure 91 – Disaggregation results of the Air Compressor using Sequence-to-point.
- Figure 92 – Disaggregation results of the CROUS using Sequence-to-point.
- Figure 93 – Disaggregation results of the TD-GF using Sequence-to-point.
- Figure 94 – Disaggregation results of the “Fantôme” load using Sequence-to-point.
- Figure 95 – Flowchart process of a standard energy audit.
- Figure 96 – Potential savings estimation with the repair of compressed air leakage.
- Figure 97 – Example of association between power and flow rate of a rotary-screw fixed speed air compressor.
- Figure 98 – Power data for the Air Compressor.
- Figure 99 – Seven days rolling average of air compressor power and flow rate.
- Figure 100 – Power data during the training period.
- Figure 101 – Global consumption load curve in a no-production period.
- Figure 102 – Power data during the training period.
- Figure 103 – Compressed air production scheme.
- Figure 104 – Power data for the air compressor identifying the power associated to both load and unload states.
- Figure 105 – Power data for the Air Compressor.
- Figure 106 – Power data for the global consumption.
- Figure 107 – Power data for the Crous and AHU loads.
- Figure 108 – Power data for the TD-GF, ASI and Fantome loads.
- Figure 109 – Disaggregation results.
- Figure 110 – Direct power measurement of the air compressor.
- Figure 111 – Power data for the Air Compressor.
- Figure 112 – Zoom in in the power data for the Air Compressor.
- Figure 113 – Power data for the global consumption.
- Figure 114 – Power data for the Crous and AHU loads.
- Figure 115 – Power data for the TD-GF, ASI and Fantome loads.
- Figure 116 – Zoom in on disaggregation results.
- Figure 117 – Annual savings with the repair of compressed air leaks.



## List of Tables

- Table 1 – Examples of existing datasets of energy consumption in buildings.
- Table 2 – Number of files accessed by each notebook.
- Table 3 – Average power according to building’s occupancy.
- Table 4 – Average power of TGBT1 and TGBT2.
- Table 5 – Average power according of TGBT1 main loads.
- Table 6 – Average power of TGBT2 main loads.
- Table 7 – Average power of TGBT1 according to occupancy periods.
- Table 8 – Average power of TGBT2 according to occupancy periods.
- Table 9 – Annual compressed air system electricity consumption considering the shutdown of the system during non-occupancy periods.
- Table 10 – Annual compressed air system electricity consumption considering the shutdown of the system during non-occupancy periods.
- Table 11 – Data quality dimensions. [7]
- Table 12 – Example of comparison between real and virtual meters of GreEn-ER.
- Table 13 – Comparison between real and virtual meters considering the average power over one year.
- Table 14 – TGBT2's virtual meter accuracy as a function of the sampling interval and the tolerance allowed.
- Table 15 – Completeness due missing samples.
- Table 16 – Outliers inserted in the data series.
- Table 17 – Performance of the regression methods on the adapted data.
- Table 18 – Number of outliers found in the global search by each method on adapted data.
- Table 19 – Number of outliers found by the Forecast Error method applied to the random forest forecasts on the adapted data.
- Table 20 – Number of outliers on real data found manually.
- Table 21 – Performance of the regression methods on the Real Data.
- Table 22 – Outliers found in the global search by each method on real data.
- Table 23 – Number of outliers found by the forecast error method applied to the random forest forecasts on the real data.
- Table 24 - Average Power of the target loads.
- Table 25 – Combinatorial optimization disaggregation results.
- Table 26 – FHMM disaggregation results.
- Table 27 – LSTM-based network disaggregation results.
- Table 28 – Window GRU network disaggregation results.
- Table 29 – Sequence to point disaggregation results.
- Table 30 – Manufacturer performance table of a Fixed-speed air compressor [14].
- Table 31 – Manufacturer performance table of a Fixed-speed air compressor. [19]
- Table 32 – Average air compressor power and flow rate during the different operation periods
- Table 33 – Average air compressor power and flow rate during the different operation periods
- Table 34 – Average power of the loads during the training period.
- Table 35 – Average air compressor power and flow rate during a no-production period.
- Table 36 – Average air compressor flow rate after the repair of the leaks.
- Table 37 – Average air compressor power after the repair of the leaks.
- Table 38 – Annual electricity savings with the repair of the leaks.
- Table 39 – Average air compressor power and flow rate during the different operation periods
- Table 40 – Average power of the loads.
- Table 41 – Average air compressor power and flow rate during a no-production period.
- Table 42 – Average air compressor power and flow rate during the different operation periods



---

Table 43 – Average power of the loads.

Table 44 – Average air compressor power and flow rate during the different operation periods

Table 45 – Comparison between leakage estimation from measurements and from NILM.

Table 46 – Average air compressor power after the leaks repair during the normal operation period.

Table 47 – Average air compressor power after the leaks repair during the no compressed air consumption period.

Table 48 – Annual savings with the repair of compressed air leaks.

## Acronyms

- Air Handling Unit – AHU
- Artificial Intelligence – AI
- Artificial Neural Network – ANN
- Auto Regressive Moving Average – ARMA
- Building Management System – BMS
- California Energy Commission’s Public Interest Energy Research – PIER
- CNN – Convolutional Neural Network
- Comma Separated Values – CSV
- Data Quality – DQ
- Digital Object Identifier – D.O.I
- Domestic Hot Water – DHW
- Electric Power Research Institute – EPRI
- European Union – EU
- Factorial Hidden Markov Model – FHMM
- Gated Recurrent Unit – GRU
- GreEn-ER – *Grenoble - Energie – Enseignement – Recherche*
- Greenhouse Effect Gas – GHG
- Heating Ventilation And Air Conditioning – HVAC
- Hidden Markov Model – HMM
- Hierarchical Data Format Version 5 – HDF5
- International Energy Agency – IEA
- Internet Of Things – IoT
- Interquartile Range – IQR
- Light-Emitting Diode – LED
- Long Short Term Memory – LSTM
- Machine Learning – ML
- Mean Absolute Deviation MAD
- Mean Absolute Error – MAE
- Mean Absolute Percentage Error – MAPE
- Mean Squared Error – MSE
- Medcouple – *MC*
- *Ministère De La Transition Ecologique* – MTE
- Nearly Zero-Energy Building – Nzeb
- NILM – Non-Intrusive Load Monitoring
- Recurrent Neural Network – RNN
- Standard Deviation – STD
- *Tableau General De Basse Tension* – TGBT



## **Introduction**

Buildings, either residential or tertiary, have a key role in the energy transition, as they represent an important share of both global energy consumption and greenhouse gas emissions. Regarding that, energy audits are powerful tools that help the identification and the quantification of potential savings in these environments. Thus, the development of methods, models and tools to enhance energy audits in tertiary buildings is one of the objectives of this thesis. Such objectives involve the comprehensive analysis of data from the GreEn-Er building at Grenoble-INP, offering an extremely rich and promising field for experimentation and extrapolation. The building is massively monitored with over 1500 sensors, of which over 300 are electricity meters.

However, because of the limited time that the auditors have to collect data from on-site measurements or the lack of historical data, it can be said that standard energy audits take a snapshot of the current conditions of a facility. In this manner, some operating modes, which can hide saving potentials, may not be addressed by the audit. In the context of energy audits, Non-Intrusive Load Monitoring (NILM) methods can be useful for enhancing the analysis performed, retrieving the load curve of unmonitored operation modes. However, large part of the studies applying NILM techniques in buildings regards to the residential sector. Hence, another objective of this thesis is to evaluate the application of energy disaggregation techniques in tertiary environments, and their role in the enhancement of energy audits in this type of building.

As mentioned in the previous paragraph, studies regarding the electricity consumption in the residential sector are numerous, much as a result of the wide availability of open-format building consumption data for this sector. In contrast, datasets for tertiary buildings are rare. Therefore, making GreEn-ER datasets openly available to promote the advancement of research in non-residential environments is also one of the goals of this thesis.

In addition, real consumption data usually intrinsically brings data quality problems, either due to acquisition or storage failures. There are several data quality problems that can be found in a dataset such as the GreEn-ER one. Among them are lack of completeness and the presence of outliers. The testing and development of methodologies to improve data quality, detecting and possibly correcting outliers in the energy consumption data is also one of the objectives of this thesis.

The present thesis is divided into eight chapters. Every chapter presents its own reference list. The first presents the scientific context in which the work is inserted,

highlighting the roles of the buildings in the energy transition in Europe, specially the tertiary sector.

The second chapter is dedicated to address the need of real consumption datasets to the advancement of the research and how the availability of residential datasets made possible the expansion of studies in this sector, leaving a lack of studies regarding tertiary buildings. Also, it presents the electricity consumption dataset of the GreEn-ER building, set up to help the progress of the research in the tertiary environment.

The third chapter discusses data quality problems and their impact in the research, presenting the dimensions associated to these problems. It also quantifies the data quality of the assembled dataset of the GreEn-ER building and presents a brief study case regarding the difference between using data with quality problems and pre-treated data.

The fourth chapter presents a new technique to detect outliers. As local outliers are difficult to detect using classic statistical methods, a hybrid one, combining prediction and a statistical method was developed. This chapter, firstly presents statistical methods for outlier detection and the Random Forest algorithm as a regression technique. Then, the combination between the predictions and the statistical methods are applied in a dataset with artificial outliers inserted. Finally the technique is tested also in real data.

The fifth chapter is dedicated to the energy disaggregation. How the disaggregation can be used to improve energy efficiency and sobriety, exposes the different state of art methods and presents the NILMTK and how its development and availability made possible the progress of the energy disaggregation in the residential sector.

The application of different disaggregation techniques to GreEn-ER building loads is presented in the sixth chapter. The NILMTK is used as framework to perform the energy disaggregation. However, to use this framework, the integration of the GreEn-ER dataset to the NILMTK is necessary. How this integration was made and the results obtained are presented in this chapter.

The seventh chapter addresses the possibility of using energy disaggregation to enhance energy audits in a tertiary building. This is discussed by using the study case of the quantification of compressed air leakage in the GreEn-ER as an example. Finally, the eighth chapter presents the final conclusions and some perspectives of future work.

Jupyter Notebooks to back up every chapter are available in the open repository found in <https://gricad-gitlab.univ-grenoble-alpes.fr/martgust/Worksheets> [These](#)

# 1 Scientific Context – Electricity Consumption in Buildings

This first chapter is dedicated to contextualize the thesis. It presents an overview of the energy consumption and its pros and cons over the years. In addition, buildings are responsible for an important share of the global energy consumption. Buildings, either residential or tertiary, account for almost half of the electricity consumption worldwide. Considering other types of energy, the GHG (Greenhouse Effect Gases) emissions due to buildings' activities correspond to 17.2% of the total emissions all over the world. These figures bring attention to the energy consumption in buildings, especially the ones of the tertiary sector, a field whose scientific research potential has not yet been fully exploited. Hence, this chapter also exposes an overview of the electricity consumption in buildings, with emphasis on the case of France.

## **Chapter Contents**

<b>1.1 Overview of the Worldwide Energy Consumption and GHG emissions</b> -----	<b>22</b>
<b>1.2 Electricity Consumption in Buildings</b> -----	<b>27</b>
<b>1.3 Conclusions</b> -----	<b>34</b>

## 1.1 Overview of the Worldwide Energy Consumption and GHG emissions

The greatest challenge of our time is to combat anthropogenic climate change. The Industrial Revolution, at the end of the 18th century, allowed a great technological advance, having a significant impact on people's quality of life. Life expectancy at birth grew significantly, especially considering the Second Industrial Revolution, with the advent of electricity, from the second half of the nineteenth century and the beginning of the twentieth century, as shown in Figure 1.

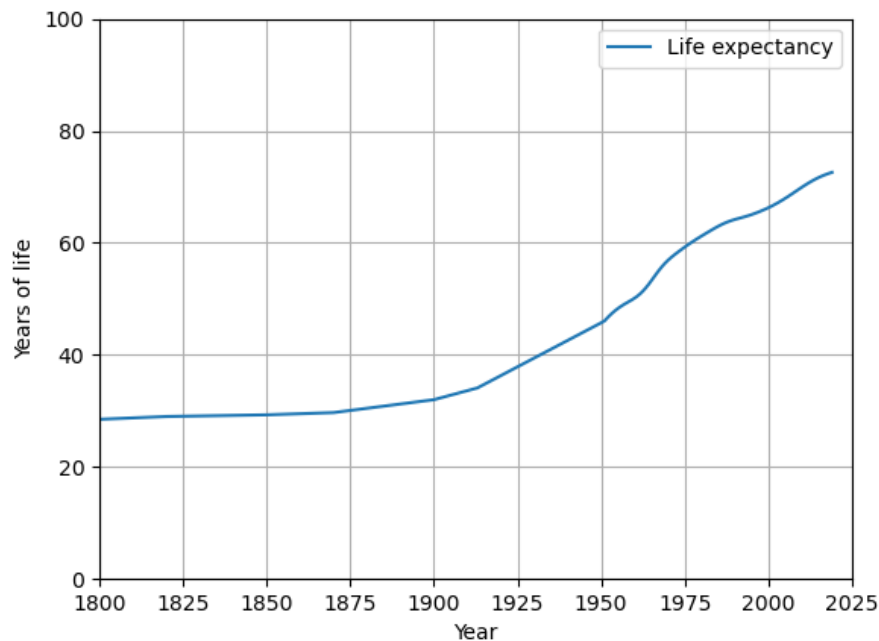


Figure 1 – Evolution of the worldwide life expectancy at birth. [1]

However, this increase in people's quality of life did not come at zero cost. The technological advance was achieved thanks to the burning of fossil fuels, such as coal initially. The burning of fossil fuels, intensified in the 20th century, is largely responsible for the greenhouse effect gas (GHG) emissions, along with agriculture, land use and forestry. As an example, energy related emissions corresponded to 73% of the total GHG emissions in 2016, as can be seen in Figure 2.

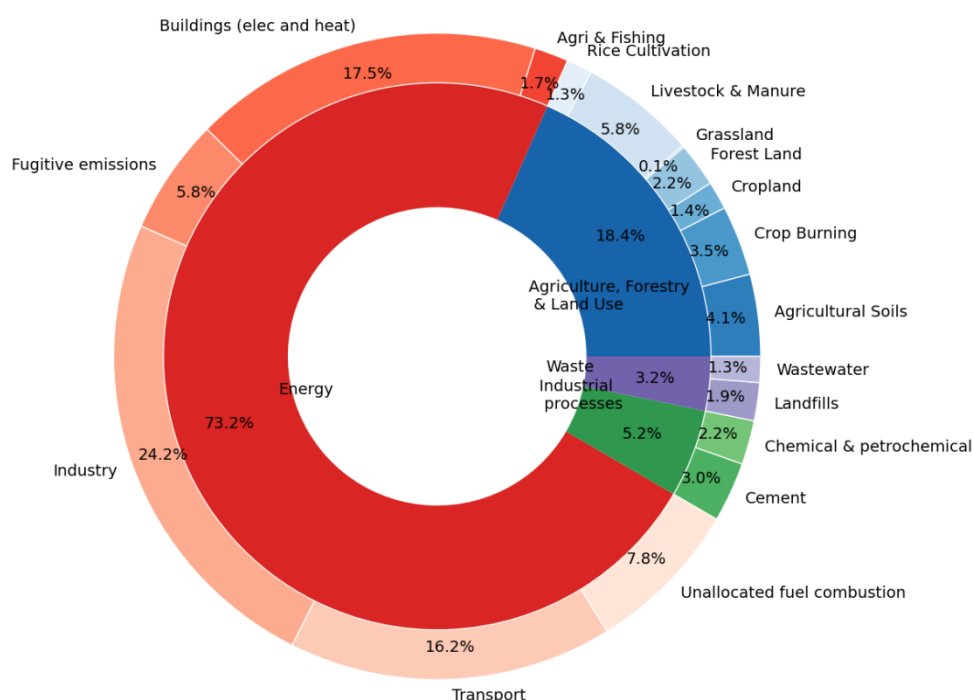


Figure 2 – Global GHG emissions by sector. [2]

The Greenhouse effect is the main responsible for the climate change. Some estimates indicate that, in 2017, human-induced warming reached approximately 1°C, regarding pre-industrial levels (1850s to 1900s) [3]. Some projections indicate that the global warming of 2°C by 2100 would be catastrophic, as the sea level would rise about 25cm [4]. Because of that, some efforts were made by the scientific community and governments, materialized in the 2015 Paris agreement, in order to limit the global warming in 1.5°C by 2100, what would reduce the sea level rise to 19cm. As an example of how fast the global temperature is increasing, in the decade 2006–2015, warming reached 0.87°C relative to 1850–1900, predominantly due to human activity increasing the amount of greenhouse gases in the atmosphere. Given that global temperature is currently rising by 0.2°C per decade, human-induced warming reached 1°C above pre-industrial levels around 2017 and, if this pace of warming continues, would reach 1.5°C around 2040.

Since energy consumption is one of the main causes of GHG emissions, which in turn is a major contributor to climate change, it is clear that an energy transition from the use of fossil fuels to renewable energy sources is necessary. However, some of the most widespread renewable sources used today, such as solar and wind power, are intermittent, making it difficult to control the stability of the power grid. Another key point to the reduction of GHG



emissions, and thus in the energy transition, is the energy management through the concepts of energy flexibility, energy efficiency and energy sobriety.

There are several kinds of energy used nowadays, such as electricity and heat for example. However, the best energy is that which we do not consume. That is the concept of energy sobriety, do not consume the energy when it is not needed. On the other hand, energy efficiency is providing the same service while consuming less energy. Finally, energy flexibility is linked to the system's ability to time shift its energy consumption according to the production from renewable sources in order to privilege these sources.

The Directive 2012/27/EU [5] of the European Council pointed to the need to increase energy efficiency in the European Union (EU) in order to achieve the target of saving 20% of primary energy consumption in the region compared to projections for 2020. The Council indicated as the main strategy the exploitation of the considerable potential for reducing energy consumption in buildings, transport and production processes. As the targets set by the 2012 directive were not on the verge of being achieved in the EU as a whole, the European Council decided to postpone the reduction targets to the year 2030 through the Directive 2018/2002 [6], but increasing from 20% in 2020 to 32.5% in 2030.

In this context, France is one of the countries where the target is furthest from being reached. For example, in 2019, primary energy consumption was still 7% above the target for that year. In Germany, for example, consumption was only 2.2% above the target, while in Italy and in the United Kingdom the consumption target for the year 2019 had been reached [7]. The primary energy consumption and the target value for 2019 in Germany, France, The United Kingdom and Italy can be seen in Figure 3.

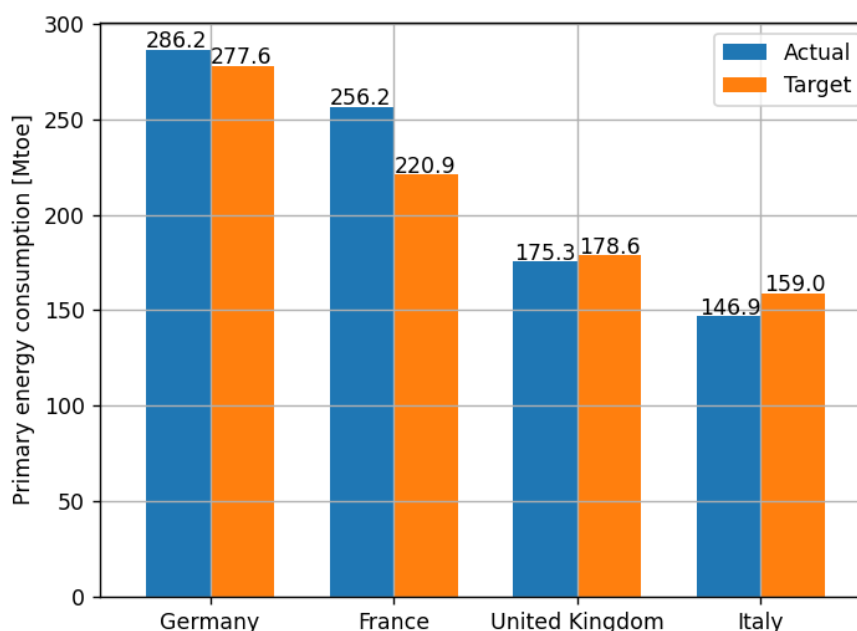
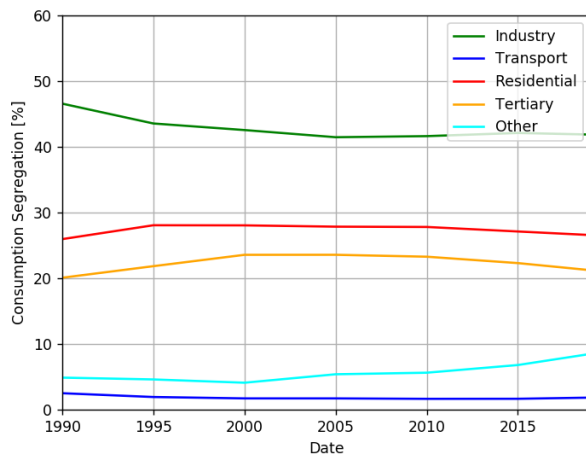
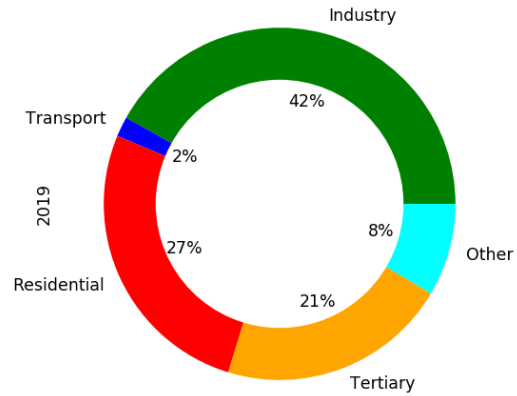


Figure 3 – Annual primary energy consumption in 2019 in selected countries of the EU.

In the challenges of energy transition and reducing consumption, buildings play a key role, as recognized by the European Council itself. Worldwide, the energy consumed in buildings represents a significant part of global energy consumption. Considering electrical energy alone, according to the International Energy Agency (IEA) [8], in 2019 about 50% were consumed in residential or tertiary buildings, slightly above the industrial consumption (42%). From 1990 until 2005, the share of industrial consumption decreased from 47% to 41%, with stabilized values to nowadays. In residential buildings, the share reached the peak between the years 1995 and 2000, with 28% before stabilizing in 27% until this day. In 1990, the share of the tertiary sector in the worldwide electricity consumption was about 20%. This portion slightly increase in the following years, reaching about 24% between 2000 and 2010 before dropping to 21% in 2019. The evolution of the worldwide electricity consumption shares by sector from 1990 to 2019 in Figure 4a, while Figure 4b details the year of 2019.



a) Evolution of the distribution by sector

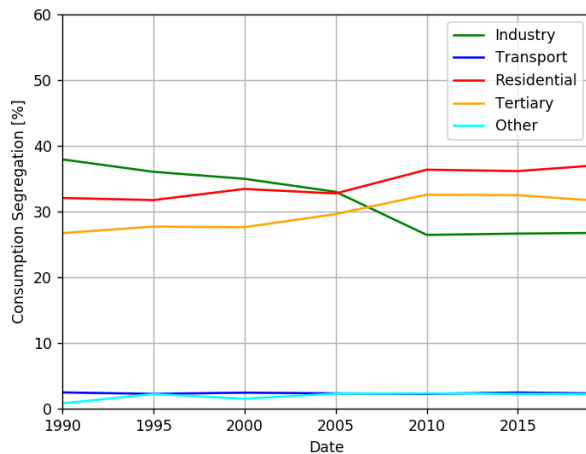


b) 2019 distribution by sector

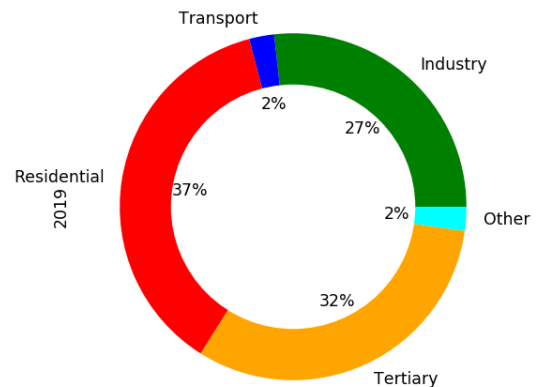
Figure 4 – Distribution of worldwide electricity consumption by sector.

Following developed countries tendency, the share of electricity consumption in buildings in France, whether residential or tertiary, is even higher. While the industrial sector saw a drop of its share, from 37% to 27% in 30 years span (from 1990 to 2019) the residential and tertiary sectors increased their importance. The share of electricity consumption of the residential sector increased 5% during the same period. Similar growth happened in the tertiary sector. The evolution of the electricity consumption in France, by sector, can be seen in Figure 5a [8].

The residential and tertiary sectors combined are responsible for nearly 70% of the country's total electricity consumption. As can be visualized in Figure 5b, the residential sector is responsible for 37% of consumption, while tertiary buildings represent for 32% of the total French electricity consumption [8].



a) Evolution of the distribution by sector



b) 2019 distribution by sector

Figure 5 – Distribution of electricity consumption by sector in France.

The great share of both residential and tertiary sectors shows the importance of these sectors in electricity consumption in France and indicates the great potential for energy savings in buildings. The next section details the typical electricity consumption in both sectors.

## 1.2 Electricity Consumption in Buildings

The electricity consumption in the residential and the tertiary buildings are very different. In terms of the global consumption, while the tertiary buildings concentrate its consumption during the day of the weekdays, energy consumption in residential buildings tends to be concentrated at evenings and on weekends. In addition, the nature of the loads is very different regarding these two environments. Furthermore, even among the tertiary buildings, the nature of the loads differs according to their activity, as the loads in a hospital, a university, an office building, a retail store or in a restaurant are very different. Because of that, the next sections explore the distribution by end use of the electricity consumption in the residential and in the tertiary sector.

### 1.2.1 Residential Buildings

As mentioned in the previous section, the residential sector represented, in 2019, 37% of the electricity consumption in France. This means that residential buildings are the major electricity consumers in the country. Because of that, studies regarding the energy consumption in this type of building are numerous.

Despite having similar general appliances, each household is different. Different sizes, heated area, number of inhabitants, people's habits, type and number of appliances and their efficiency significantly influence the energy consumption of a household. Because of that, the residential consumption is modelled considering eight major usages [9]:

- Heating,
- Domestic hot water (DHW),
- Ventilation and air conditioning,
- White goods (refrigerator, washer etc.),
- Audiovisual and informatics (TVs, personal computers etc.),
- Cooking,

- Lighting,
- Other usages.

Figure 6 shows the distribution of consumption by use in the French residential sector in 2019 [9].

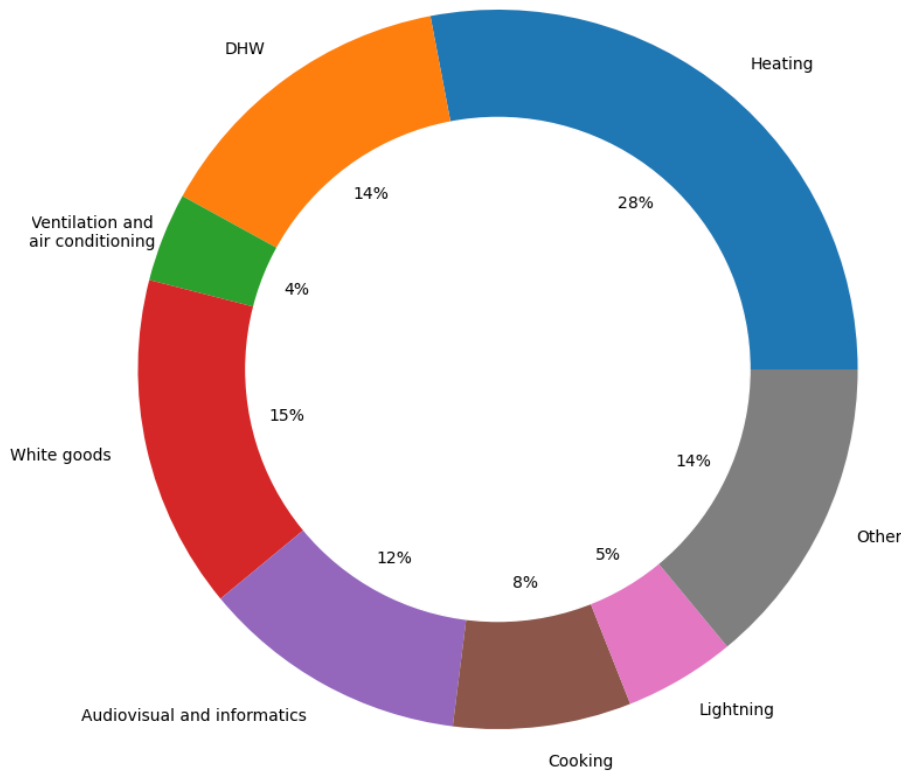


Figure 6 – Distribution of consumption by use in the French residential sector in 2019.

According to the Bilan RTE 2021 [9], as a result of the electrification of the building stock, heating consumption grew strongly until the early 2010s, with an average annual growth rate of 3.2% per year between 2005 and 2012. However, the introduction of the 2012 thermal regulation, which favored gas installations in new construction, and the improvement of the energy performance of buildings resulted in a strong slowdown in the growth of electric heating consumption, with an average change of around 0.8% per year between 2012 and 2019. The new environmental regulation, the RE2020 [11], which was initially presented in 2020 and then modified in 2021, takes more into account environmental constraints, particularly on CO<sub>2</sub> emissions. These thresholds are almost equivalent to a ban on fossil fuel heating sources in new construction, and could thus contribute to increase the market share of electric heating solutions in the future. In France, electric heating represents 28% of residential electrical consumption in 2019 and tends to rise in the following years. Another important consumption is due to ventilation and air conditioning, as it accounted for 4% of the

French residential sector's consumption in 2019. If we take only the air conditioning, it alone represents 2.4% of total consumption in 2019. This share also tends to rise as the years are becoming hotter due to the climate change.

Similar analysis can be done for the water heating. This consumption grew significantly in the early 2000s, by around 2% per year between 2005 and 2012, due to the electrification of the stock, before undergoing a phase of strong slowdown (an average of 0.2% per year between 2012 and 2019), under the effect of the 2012 thermal regulations in a similar way to the evolution observed for heating. However, the introduction of CO<sub>2</sub> emission thresholds by the RE2020 [11] limits the use of gas solutions for the water heating, prioritizing the use of electricity in this task. Thus, the market share of electricity for the production of domestic hot water in new construction should evolve in a similar way to that of the market share for heating, as the solutions are often coupled. In 2019, domestic hot water (DHW) production represented 14% of the French residential sector's electricity consumption.

White goods, which include refrigeration and washing appliances, whose consumption represents 15% of residential consumption in 2019 in French metropolitan territory, are gradually being replaced by less energy-intensive appliances. This dynamic is being driven by the implementation of ambitious European Union directives on eco-design and energy labelling, which are helping to reduce consumption despite rising demographic trends.

The digital revolution led in the 2000s to the massive arrival in homes of computer and internet equipment. Arriving in 2019, all computer and audio-visual uses represented were responsible for 12% of the consumption of the French residential sector. While the number of electrical equipment in this item has risen sharply in recent years, they are undergoing a major transformation with a rationalization of uses and a change in lifestyles. Older fixed devices are gradually being replaced by less energy-intensive mobile devices, for example fixed computers are being replaced by laptops and even digital tablets. Indeed, in 2019, smartphones were preferred by 51% of French people and tablets by 12% for internet access, while only 31% of French people preferred the computer. Therefore, even if the number of this kind of equipment tends to rise, the rise in the consumption does not follow this trend at the same rate.

In France, cooking use accounted for 8% of residential electricity consumption in 2019 as it is steadily gaining market share at the expense of gas cooking. The advent of induction cooking, which combines the safety of electricity with cooking comfort similar to that of gas, is driving the electrification of this use. Electric hobs accounted for nearly 80% of cooktop sales in 2019. This electrification, combined with an increase in other cooking

equipment such as microwave ovens, is being offset by significant progress in energy efficiency, particularly for ovens, which are subject to European directives on eco-design and energy labelling, and for electric hobs, as induction hobs consume less energy than glass-ceramic hobs, which themselves consume less energy than cast iron hobs.

Recent European directives have banned, since late 2018, the sale of high energy consuming lamps, including incandescent and halogen lamps with the exception of certain specific models that have no alternative to date. The arrival on the market of LED (Light-emitting diode) bulbs at competitive prices since the mid-2010s causes a drastic reduction in electricity consumption in lighting, they even accounted for three quarters of lamps sold for residential in 2017. As presented in Figure 6, in 2019, the lightning consumption represented 5% of residential consumption in France.

### 1.2.2 Tertiary Buildings

The tertiary sector covers a wide range of activities that can be classified as mainly commercial (trade, transport, financial activities, services to businesses, services to individuals, lodging and catering, real estate, information and communication) and non-commercial (public administration, education, human health, social activities etc.). This sector represented nearly 32% of the electricity demand in metropolitan France. [8]

The electricity consumption of the tertiary sector can be divided in two main partitions. The first one, the in-building branch comprehends all loads located inside buildings, which represents 81% of the whole consumption [9]. The other one, called off-building branch, comprehends activities in which consumption is mainly linked to an industrial process, like telecommunications, waste treatment and consumptions that are located literary outside a building, like the public lightning and accounts for the other 19% of the electricity consumption of the tertiary sector [9].

As previously mentioned, the tertiary sector is a very heterogeneous one, including activities such as health, commerce, cultural venues or offices. In order to describe the consumption of the building sector, activities with a similar consumption profile are grouped together in eight branches of activity [9], enumerated as follows:

- Cafés, hotels, restaurants (Catering)
- Community housing
- Health care
- Education

- Sports, culture, leisure, various community facilities
- Offices and administration
- Commerce
- Transportation

The electricity consumption inside buildings, in France during 2019, by activity can be seen in Figure 7 [9].

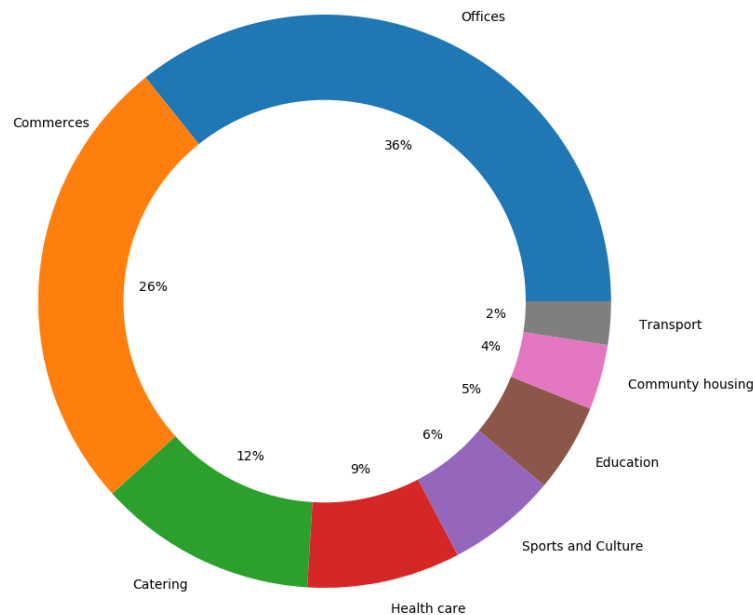


Figure 7 – Distribution of consumption by activity of tertiary buildings in France in 2019.

As seen in the previous figure, commerce and offices represent, more than 60% of the in-building electricity consumption in the tertiary sector. While restaurants, hospitals and educational buildings represent 12%, 9% and 5% of the consumption of tertiary buildings.

For each of the eight branches of activity of the in-building sector, the consumption of seven usages is modeled [9]: heating, air conditioning, domestic hot water, cooking, lighting, cooling and specific electricity. The distribution of the electricity consumption by type of appliance in tertiary buildings in France in 2019 is illustrated in Figure 8.



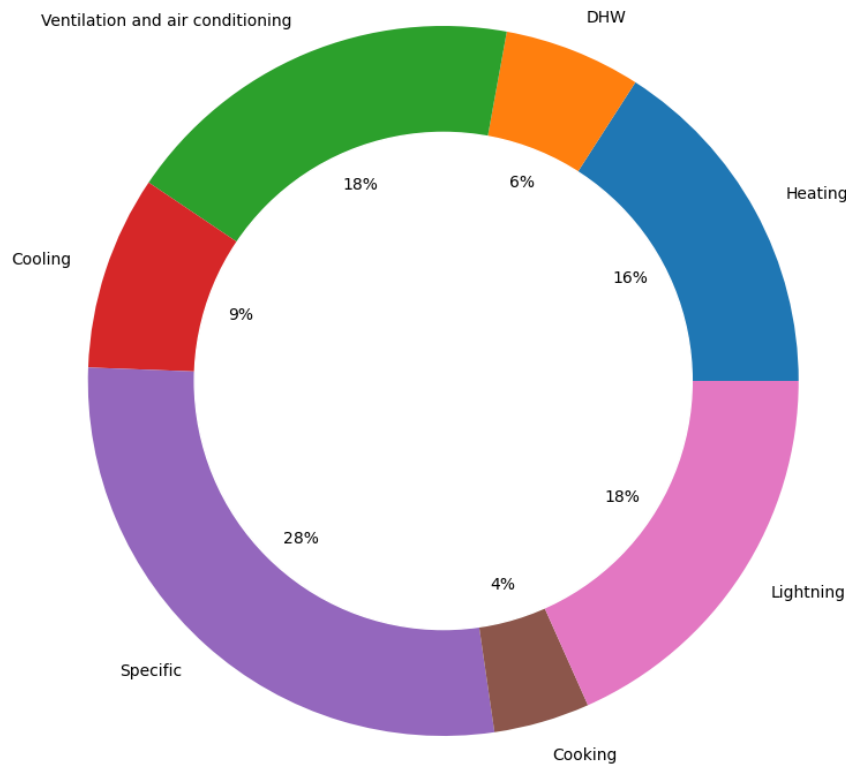


Figure 8 – Distribution of consumption by appliance type of tertiary buildings in France in 2019.

Similarly to residential buildings, the heating loads go through an electrification process. In 2019, the market share of electric heating in newly constructed buildings was estimated at 53%. Also, a significant part of the surface areas built are offices and commerce, which make extensive use of electric heat pumps, that can also be used for air conditioning taking advantage of their reversibility. In the same year, electric heating represented 16% of the electricity consumption in tertiary buildings, being the fourth major consumer inside this kind of construction [9].

The production of domestic hot water (DHW) represented, in 2019, 6% of the consumption of the tertiary buildings. This electrical consumption has increased by 14% since 2005, with an inflection of this growth since 2012, similar to what happened in the residential sector. This inflection is mainly due to efficiency gains related to the systems, to the installation of thermodynamic water heaters instead of Joule water heaters [9].

Air conditioning consumption has sharply risen. Since 2005, it has increased by more than 60% in proportion to the increase in air-conditioned surfaces. However, this growth has slowed since 2012. While, between 2005 and 2012, the average annual growth rate was 5%, it has dropped to 1.5% between 2012 and 2019. While the Education and Community Housing branches are not cooled, at 5-10%, nearly half of the surfaces in the Offices, Administrations

and Cafés, Hotels, Restaurants branches are air-conditioned. Because of this, the ventilation and air conditioning already represent 18% of the electricity consumption of the tertiary buildings, in France.

Electricity consumption for cooking purposes is currently 4% of the whole consumption tertiary buildings. This consumption mainly covers cooking equipment used in restaurants, whether in company restaurants, collective catering or commercial restaurants. Cooking is also characterized by an electrification of the use, as a consequence of an evolution of the practices which tend to be favorable to electricity. Approximately 70% of the newly built surfaces are equipped with electric cooking systems. Installed in similar environments, the cooling loads represent around 9% of the electricity consumption of this sector [9].

Indoor lighting represents a significant potential for energy efficiency in the tertiary sector. The widespread use of LED technology has recently been encouraged by the sale of lamps that are compatible with fluorescent systems that previously required a heavy replacement operation. In 2016, less than 10% of offices were equipped with LED systems in France [9]. At the same time, policies to eliminate the oldest technologies are forcing players to turn to the most efficient technologies. The widespread use of LED lamps is expected in the long term, even in developing countries. Meanwhile, in 2019, the indoor lightning consumption represented 18% of the whole tertiary sector consumption in France [9].

Other specific electric uses include all uses of electricity not already described above. These include office equipment, like personal computers, printers, data centers, medical equipment, elevators and escalators. There are also some household appliances, such as vacuum cleaners, coffee machines, cold drink dispensers, etc. As in the residential sector, the spread of technical progress is expected to continue under the effect of European ecodesign regulations that impose increasingly stringent standards for energy efficiency. The consumption of this equipment is estimated at 28% of the whole electricity consumption of the French tertiary buildings.

### **1.3 Conclusions**

This first chapter was dedicated to contextualize the thesis. It has presented the important share of the energy consumption in buildings. For instance, the residential and tertiary sectors combined are responsible for nearly 70% of total electricity consumption in France, being the residential sector is responsible for 37% of consumption, while tertiary buildings represent for 32% of the total French electricity consumption. These figures bring attention to the energy consumption in buildings, especially the ones of the tertiary sector, a field whose scientific research potential has not yet been fully exploited.

## References

- [1] Roser, M. Life Expectancy. Our World in Data, 2019, Retrieved April 7, 2022, from <https://ourworldindata.org/life-expectancy>
- [2] Ritchie, H. Emissions by sector. Our World in Data, 2020, Retrieved April 7, 2022, from <https://ourworldindata.org/emissions-by-sector>
- [3] IPCC, 2018: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. In Press.  
[https://www.ipcc.ch/site/assets/uploads/sites/2/2019/05/SR15\\_Chapter1\\_Low\\_Res.pdf](https://www.ipcc.ch/site/assets/uploads/sites/2/2019/05/SR15_Chapter1_Low_Res.pdf)
- [4] Jaganmohan M. Global sea level rise contributions by 2100, by scenario. Statista, 2022, Retrieved April 7, 2022, from <https://www.statista.com/statistics/1296870/contributions-to-global-sea-level-rise/>
- [5] Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC (Text with EEA relevance) OJ L 315, 14.11.2012, p. 1–56. <http://data.europa.eu/eli/dir/2012/27/oj>
- [6] Directive (EU) 2018/2002 of the European Parliament and of the Council of 11 December 2018 amending Directive 2012/27/EU on energy efficiency (Text with EEA relevance.) PE/54/2018/REV/1 OJ L 328, 21.12.2018, p. 210–230 <http://data.europa.eu/eli/dir/2018/2002/oj>
- [7] Eurostat, Energy efficiency statistics - Statistics Explained. 2021. Retrieved April 7, 2022, from [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy\\_efficiency\\_statistics#Primary\\_energy\\_consumption\\_and\\_distance\\_to\\_2020\\_and\\_2030\\_targets](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_efficiency_statistics#Primary_energy_consumption_and_distance_to_2020_and_2030_targets)
- [8] International Energy Agency. IEA World Energy Balances database. 2021. [Online] Available: <https://www.iea.org/data-and-statistics/data-product/world-energy-balances>
- [9] Réseau de Transport d'électricité, Bilan Électrique 2018, 2019, [Online], Available: [https://www.rte-france.com/sites/default/files/be\\_pdf\\_2018v3.pdf](https://www.rte-france.com/sites/default/files/be_pdf_2018v3.pdf)
- [10] Réseau de Transport d'électricité, Bilan prévisionnel de l'équilibre offre-demande d'électricité en France, 2021, [Online] Available: <https://assets.rte-france.com/prod/public/2021-05/Bilan%20previsionnel%202021%20-%20annexes%20techniques.pdf>
- [11] Ministère de la Transition écologique (MTE). RE 2020 - Éco-construire pour le confort de tous, 2021, [Online] Available: [https://www.ecologie.gouv.fr/sites/default/files/2021.02.18\\_DP\\_RE2020\\_EcoConstruire\\_0.pdf](https://www.ecologie.gouv.fr/sites/default/files/2021.02.18_DP_RE2020_EcoConstruire_0.pdf)



## 2 Energy Consumption Datasets of Buildings – Objectives, State of Art and The Green-ER Case

Chapter 2 exposes the importance of available datasets to the advance of the research in buildings’ energy consumption field and presents some examples of available datasets to date. It can be seen that most of them concerns the residential sector, uncovering a lack of datasets regarding the tertiary segment. To address this issue, a dataset containing aggregated and disaggregated energy consumption data of the GreEn-ER building were gathered and made available in open access. Hence, the present chapter also details this dataset, its metadata and presents some insight about the building’s consumption, such as its electricity consumption sectorial distribution and some energy sobriety opportunities identified by this analysis.

### Chapter Contents

2.1 The Importance of Datasets Towards the Research Advance-----	38
2.2 Data Valorization – Towards Open-Science -----	39
2.3 The GreEn-ER Dataset -----	42
2.4 GreEn-ER Electricity Consumption -----	49
2.5 Conclusions-----	63

## 2.1 The Importance of Datasets Towards the Research Advance

The increased use of intermittent renewable energy sources, such as solar and wind, makes the use of machine learning (ML) methods combined with demand side management more and more frequent. This leads to the use of artificial intelligence techniques to predict the consumption, the generation from renewable sources and to calculate the optimal cost, for instance. These techniques use a massive amount of data to perform their task.

Regarding machine learning approaches, there are two main axes: supervised learning and unsupervised learning. The key difference between them is that one uses labeled data to help predict outcomes, while the other does not. The supervised learning approach requires well-identified data for training the algorithms. Algorithms that fall in this approach use input data to construct a model and compare its outcomes to labeled output data. This comparison allows the algorithm to correct the model overtime. This step is called training phase. With the fitted model, fresh input data are inserted into the model, outputting a prediction. They are used mainly for classification and regression problems. While classification problems try to separate data into specific categories, regression ones seek to find the relationship between a dependent variable and one, or several, independent variables. Unsupervised approaches, on the other hand, are often used to cluster unlabeled data based on their similarities or differences.

Further in this work, ML techniques are employed to try to perform predictions of electricity consumption of a tertiary building and to disaggregate, by applying Non-Intrusive Load Monitoring (NILM) [1] techniques, the main consumption down to target appliances. These tasks are often addressed using supervised techniques. The development of different algorithms goes through the use of labeled datasets, not only to test their results, but also to train the algorithms. Hence, the need of real datasets of electricity consumption increases continuously.

The progress of research related to the use of machine learning in the energy consumption field directly depends on the availability of datasets, either for training or performance testing. Even though datasets containing synthetic data have had their importance, datasets containing real data of electricity consumption provide, especially in the buildings field, further advances, despite increasing the difficulty in developing and applying algorithms.

Therefore, the use of real data of electricity consumption measurements for researchers to advance in the field of machine-learning in buildings is needed. There are several datasets publicly available, with both aggregated and disaggregated consumption.

Some of them are mentioned in the following tables.

Among the datasets mentioned earlier, most are dedicated to the residential sector, as it is shown in Table 1. Because of the different nature of the loads, highlighted in the previous chapter, it is obvious that there is a need to create datasets dedicated to buildings of the tertiary sector, since the equipment and the consumption profile are very different. Because it is massively monitored, GreEn-ER, a tertiary sector building, is a good example where the consumption data can be used to create exploitable datasets. Following sections explore the importance of the availability of data in open-access as well the GreEn-ER dataset that was made available within this work.

## 2.2 Data Valorization – Towards Open-Science

Currently, we live in the Data Age. Daily, more than five billion people interact with data, which will be six billion by 2025. Consumers will have one data interaction every 18 seconds, much because of the billions IoT (Internet of Things) devices connected all over the world, by when they are expected to create over 90ZB of data [24]. Thus, data are in the core of the scientific production.

The ease with which data sharing is carried out nowadays helps popularize scientific knowledge, allowing researchers from many places to develop theories from data originally obtained from other parts of the world. In addition, sharing scientific knowledge makes research more efficient, more visible, and less redundant. The unrestricted dissemination of research publication and data can be called open-science. It takes advantage of the digital transformation to promote open access to publications, once restricted under paid licenses, and to research data. Open-science initiatives also changes the way society sees the research, making it more transparent, strengthening its integrity. Furthermore, the facilitated accessibility may accelerate the response to current issues.

In line with that, the National Plan for the Open-science (“Plan national pour la science ouverte” in French) of the French government strongly encourages the open-science in France. This plan is divided into three main axes.



Table 1 – Examples of existing datasets of energy consumption in buildings.

Dataset	Description	Country	Sector	Reference
REDD	Total power and sub-meter data from six households. They were the first datasets with the main purpose of studying NILM methods and it is the most used for testing disaggregation algorithms.	USA	Residential	[2]
BLUED	Contains measurements of voltage and current with 12 kHz sampling of one household during one week. There are no sub-meter measurements but there is information about the switching on and off of each equipment that can be used in algorithms.	USA	Residential	[3]
Household Electricity Survey	Contains equipment measurements of 251 households. There are also total power data from 14 of these households.	United Kingdom	Residential	[4]
Tracebase	It is available energy consumption data of equipment used in residential and tertiary buildings that were measured with a commercial sensor.	Germany	Residential and tertiary	[5]
AMPds	Total power and sub-metering data of a household for two years.	Canada	Residential	[6]
iAWE	Contains 73 days of household and sub-meter power measurements in Nova Delhi.	India	Residential	[7]
BERDS	Contains power measurements (active, reactive and apparent) of a university campus, of several equipment, such as lighting, hydraulic pumps and air conditioning system.	USA	Tertiary	[8]
ACS-F1	Contains electricity consumption data of 100 typical appliances found in a household, such as cell phones, computers, refrigerators and TVs.	Switzerland	Residential	[9]
UK-DALE	Contains measurements of total power and some equipment of five households.	United Kingdom	Residential	[10]
ECO	Provides aggregate consumption data with one second sampling. There are also consumption data for a few selected appliances and occupancy information.	Switzerland	Residential	[11]
GREEND	Contains measured equipment from nine households in Austria and Italy, with 1-second sampling.	Austria and Italy	Residential	[12]
SustData	Contain several measurements of electrical power (active, reactive and apparent) of 50 households with a sampling time of one minute.	Portugal	Residential	[13]
COMBED	There is data from 200 meters installed in a university campus in India.	India	Tertiary	[14]
PLAID	Contains voltage and current measurements of 11 different appliances in 55 homes with 30 kHz sampling.	USA	Residential	[15]
DRED	A dataset with measurements of energy consumption of devices and information about occupancy and climate of a household.	Netherlands	Residential	[16]
Dataport	There are total and equipment consumption data for 722 households in the United States with one minute sampling.	USA	Residential	[17]
COOLL	There are voltage and current measurements of 42 different equipment with 100 kHz sampling.	France	Residential	[18]
WHITED	Contains measurements of 110 different equipment from the first five seconds of operation when it is turned on. These measurements were made with 44 kHz sampling.	World wide	Residential and small industry	[19]
REFIT	Data of total consumption of submeters of 20 households.	United Kingdom	Residential	[20]

- Generalize open access to publications.
  - Make open access publication mandatory for articles and books resulting from publicly funded research
- Be part of a sustainable, European and international dynamic.
  - Transform scientific practices to integrate open science into everyday life, so that they become a reflex and contribute to the structuring of the international open science landscape through the dissemination of best practices.
- Structure and opening up research data.
  - Make the open dissemination of research data from publicly funded programs mandatory.

### 2.2.1 Research Data Publication

Published data alone are not fully exploitable since researchers not so familiar with the dataset need to understand the data well in order to be able to leverage it in their research. Therefore, good practice indicates that published research data should be accompanied by associated metadata, some documentation, such as a data paper, and software code (in cases where raw data have been pretreated). These practices facilitate the reuse of data by different researchers

Because of the great amount of data produced, dataset publication usually is done through repositories that can guarantee that the datasets are well documented. Datasets should also be capable of being referenced in a unique and persistent manner. In that matter, repositories often deliver unique D.O.I (Digital Object Identifier) to the published datasets [26]. The Grenoble Alpes University (Université Grenoble Alpes) has its own repository, called Perscido, to host data from research performed within the university. Another example of repository is Mendeley Data, which is managed by Elsevier and is linked to a journal dedicated to data articles. As a supplementary work of this thesis, a dataset was published into two open-data repositories and will be detailed in following sections.

### 2.2.2 Data Article

A data paper is a way to well document a dataset. It can be seen as a metadata document that describes a dataset and it is published in the form of a peer-reviewed article. It comes as a complement to a dataset. This type of paper has as main objectives to make data

accessible, interpretable and reusable, rather than presenting and testing new theories, or hypotheses, or new analyses. It is important to remark that publishing a data article does not prevent the publication of traditional research articles using the data described in the data paper. Several journals are dedicated to publish data articles, often directly related to repositories. During this thesis, a data paper [21] regarding the dataset published was made available in an open science repository.

### 2.3 The GreEn-ER Dataset

The GreEn-ER building [22] is located in the Polygone Scientifique, at the Presqu'île of Grenoble, France. It gathers the Grenoble-INP engineering school Ense3, the G2Elab laboratory and training and research platforms. A photo of the building is shown in Figure 9, while Figure 10 presents a map indicating its location. The building has more than 22,000 m<sup>2</sup> of floor space, which is divided over 6 floors and the roof. There are about 1,500 students and hundreds of professors, researchers and staff using it. Because it is a large building, its power consumption is also significant. On a typical day, the active power can be more than 300 kW. There are more than 1,500 meters, including more than 300 electricity consumption ones. The electric meters measure not only the consumption of the various switchboards, regarding the aggregated consumption of different zones in the building, but also some individual loads, such as the lighting and the power outlets of certain switchboards, the air handling units (AHUs), chillers, pumps, etc. The other meters concern internal and external conditions, thermic energy data, etc. The measured data are used to control the internal conditions, regarding the comfort of the occupants and to monitor the consumption.



Figure 9 – The GreEn-ER building [22].

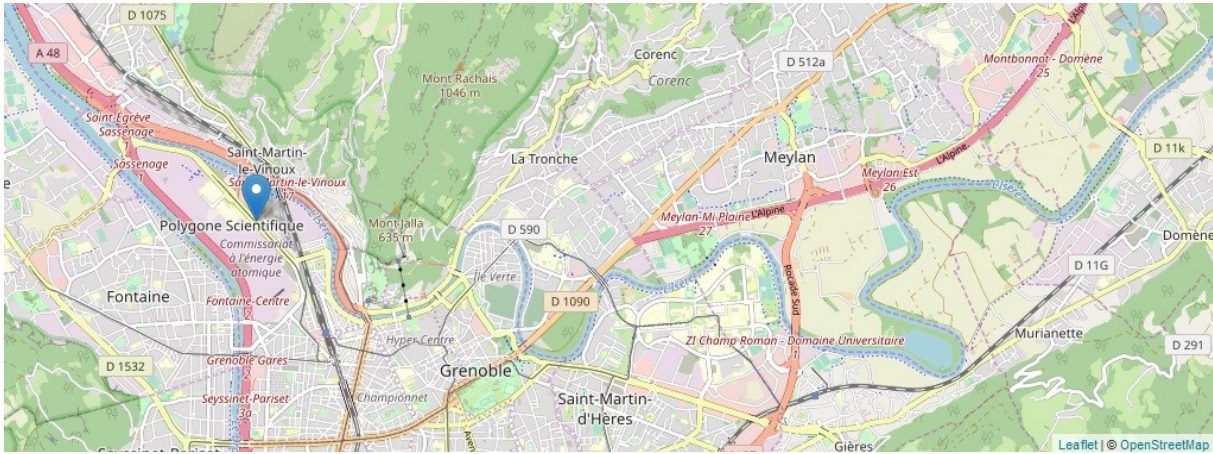
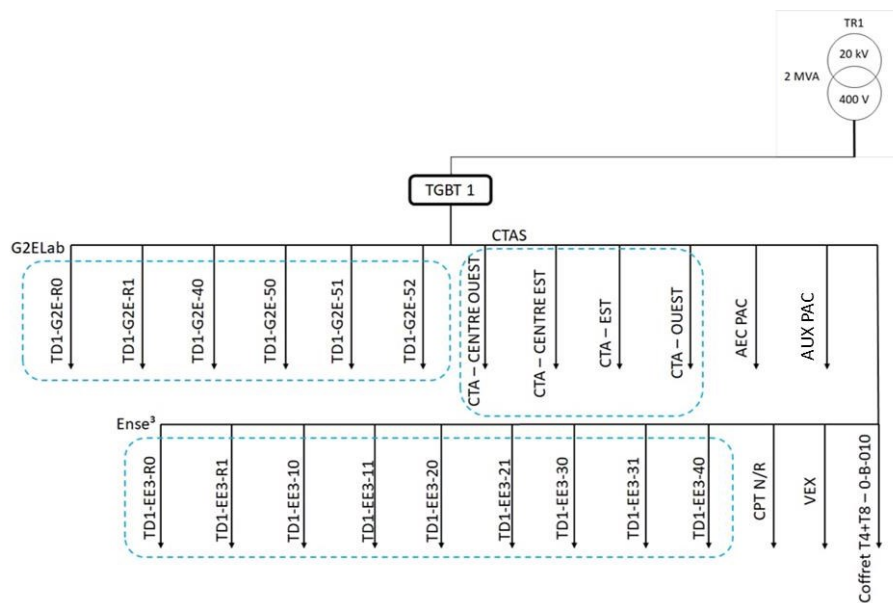


Figure 10 – Map indicating the location of the building (21 Avenue des Martyrs, 38000, Grenoble, France).

### 2.3.1 Electric Scheme

The grid delivers the electricity to the building at three-phase 20 kV. Two 2 MVA transformers (TR1 and TR2) step down the voltage to 400 V. Each transformer leads the energy to a main switchboard, called TGBT (*Tableau General de Basse Tension*), French acronym to General Low Voltage Switchboard. Each one of these boards has its own meter to measure its consumption. A switch that is normally open interconnects these boards. In that way, the two TGBTs are normally independent. Thus, all the building's loads are connected to these two main switchboards, either directly or by some sub boards. Each one of the branches of the scheme has also its own energy meter. Figure 11 illustrates the electric scheme of the building.



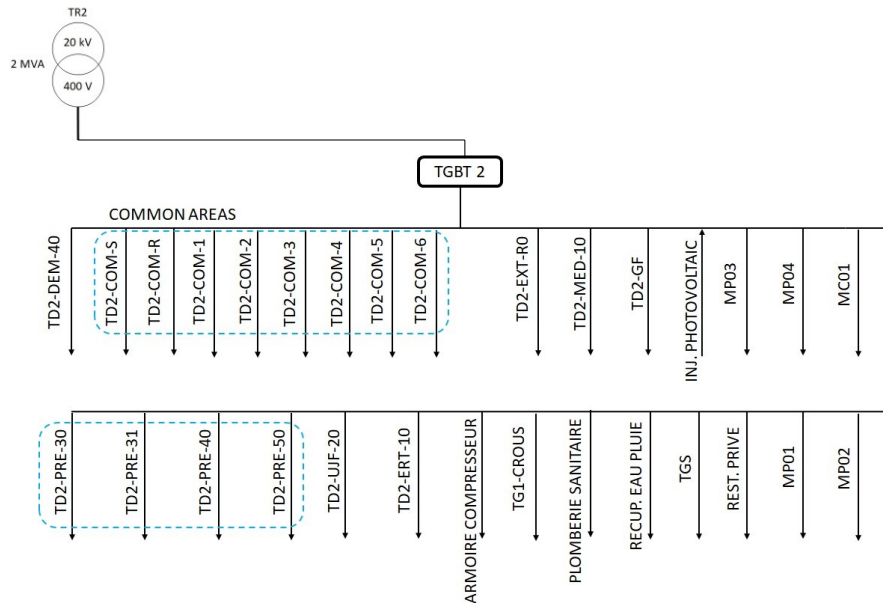


Figure 11 – Electric scheme of the GreEn-ER building.

In Figure 11, the branches that have TD in their names are, in fact, other boards that distributes the electricity to different zones. The third character in these branches' names, 1 or 2, stands for the TGBT to which the board is connected. The G2E stands for G2ELab and represents the boards that distributes electricity to that area. At the same time, EE3 stands for Ense<sup>3</sup>, and those boards distribute electricity to the classrooms and other facilities of the Engineering School Ense<sup>3</sup>. COM stands for the common areas and PRE represents the boards that distribute the electrical energy to PREDIS charges, a training and research platform for smart grids. The name's seventh character of each board is linked to the floor where it is located (R stands for Ground Floor (Rez-de-Chaussée in French)). The loads in the TD switchboards are generally divided in three or four loads, which represents lightning (ECL), outlets (PC), water heater (ECS) or dedicated outlets (FM). These loads also have their own energy meters.

Within the building, there is a platform, called “PREDIS-MHI”, conceived to be a nearly zero-energy building (Nzeb). It is a 600 m<sup>2</sup> platform energetically independent from the rest of the building. This platform, represented in the early drawing by the branch with the acronym “TD2-DEM-40”, is even more monitored than the rest of the building. In this sector, the lightning and the outlets of each room is measured independently.

### 2.3.2 Meters and Building Management System (BMS)

The electricity consumption of the building is measured by Socomec meters models E13, E23, E33, E43, E63, I30, I35 and I60, according to their specifications. Each meter has

Modbus communication via RS485 with a PLC installed in the switchboard to which the measured load is connected. The PLCs, in turn, send the measurements to the storage and to the BMS.

The BMS is based on the StruxureWare environment, developed by Schneider Electric industry. It gathers the data coming from the PLCs, and stores the measured data into a SQL server, where the data are logged. It also enables the control of some parameters, such as the internal temperature of several rooms, the air pressure and flow of the air handling units, etc. A software to manage the energy consumption, called AREE Building, developed by Inneasoftware, organizes the meters hierarchy and the trends and can show several performance indicators. It also enables the access to the logged data and can friendly export the data into text files. The data available in this dataset are extracted from the SQL server with the help of AREE Building software. The BMS can be accessed by dedicated personal computers inside the building's premises or by an on-line client. Figure 12 and Figure 13 presents screenshots from the BMS software as examples.

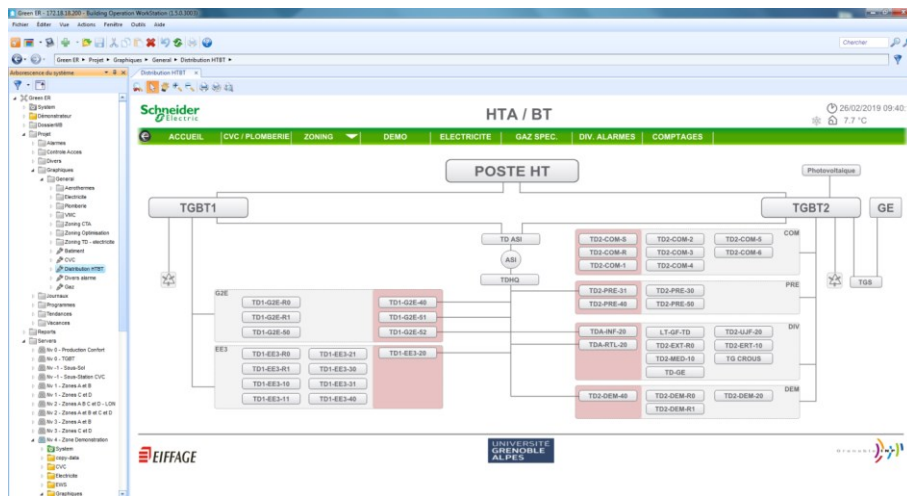


Figure 12 – Electric scheme of the GreEn-ER building as presented by the BMS.

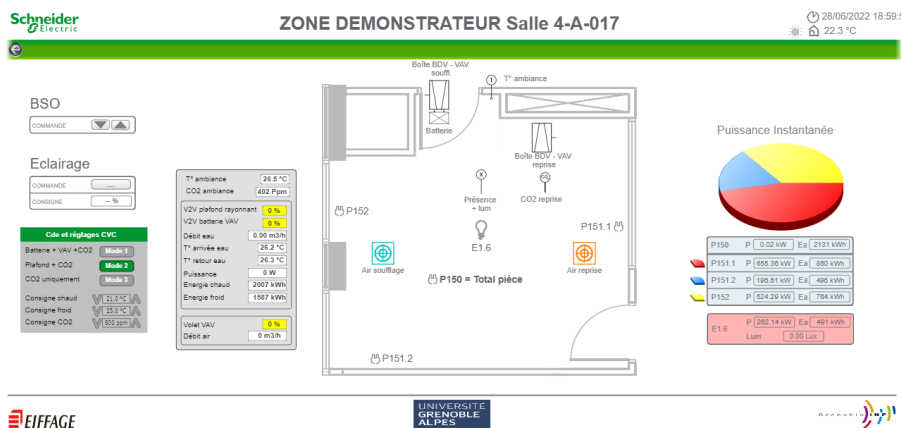


Figure 13 – On-line client of the BMS showing the parameters of the room 4-A-017, located at the fourth floor.

### 2.3.3 Dataset Structure

The dataset [23] was separated in four main contents: Global consumption, TGBT1, TGBT2 and PREDIS-MHI. In the Dataset main folder, there is a folder named “Data”. Inside this folder, two subfolders represent each year of data available, which are 2017 and 2018. Each subfolder contains three other sub-folders, and each one corresponds to a content cited earlier. Each subfolder inside "Data" also contains the CSV files with the electricity consumption data of the whole building, “585.csv” and ‘771.csv” that represents each one of the TGBTs. These folders also contain a file named "Temp.csv" with the temperature data. The temperature data are in Celsius degree °C, with one hour sampling, semicolon as separator and comma as decimal marker. The structure of folders and files is illustrated in the following figures.

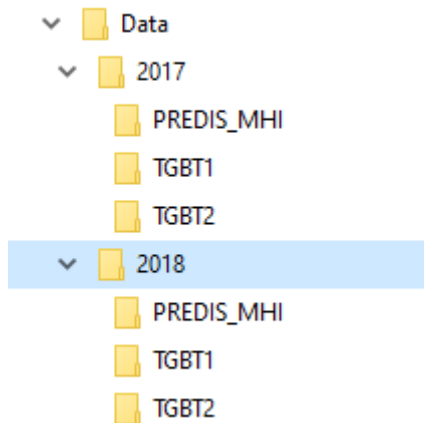


Figure 14 – Structure of the main folder of the dataset.







 PREDIS_MHI	11/06/2020 23:00	Dossier de fichiers
 TGBT1	10/06/2020 20:25	Dossier de fichiers
 TGBT2	11/06/2020 23:00	Dossier de fichiers
 585.csv	24/06/2019 14:52	Fichier CSV
 771.csv	24/06/2019 14:52	Fichier CSV
 Temp.csv	07/05/2020 18:38	Fichier CSV

Figure 15 – Structure of a folder of a specific year of the dataset.

Inside the subfolders, there are files that contain the electricity consumption data. The data are stored in CSV (comma separated values), with semicolon as separator. Each file contains the timestamp, with 10 minutes sampling and the cumulative electricity

consumption, in kWh. As the meters are cumulative and the resolution is 1 kWh, therefore the consumption sample will only increase after 1 kWh consumption of the respective load. There is one CSV file for each meter, and they are all named according to their respective meter number. These numbers, among other metadata can be retrieved in tables and drawings available in Jupyter Notebooks. Jupyter Notebook is a file format combining text, graphics and code in Python. Four notebooks are also available in the main folder, exploring all the data within the dataset.

Due to the complexity of the system and the amount of data available, four different Jupyter Notebook files were prepared. One of them explores the total consumption of the building, another one explores the TGBT1 and another one explores the TGBT2. Finally, another notebook explores the data from the PREDIS-MHI platform, a part that is energy isolated from the rest of the building. It is a living lab, gathering classrooms, offices and experimental platforms. This portion of the building also has a dedicated HVAC system, photovoltaic panels and electric vehicles stations. Figure 16 presents an overview of the PREDIS-MHI. The following table shows the amount of files that each notebook explores.

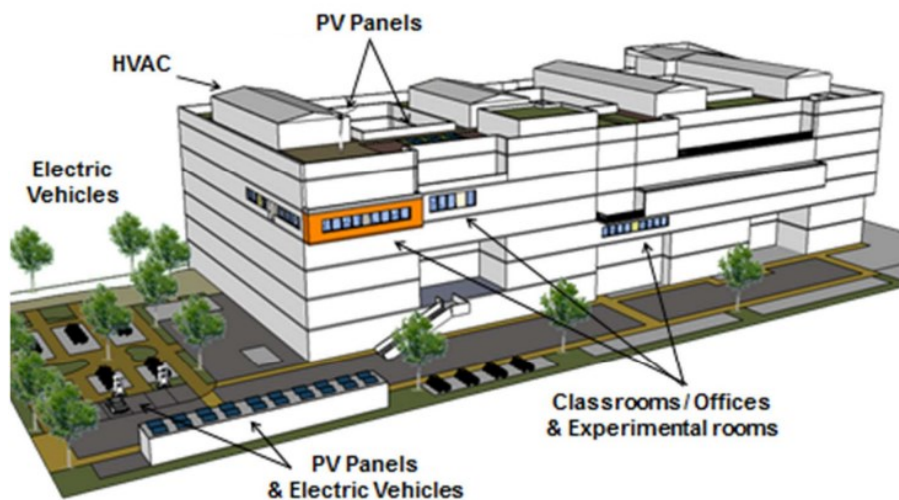


Figure 16 – Overview of the PREDIS-MHI [22].

Table 2 – Number of files accessed by each notebook.

<b>Notebook</b>	<b>Number of files</b>
GreEn-ER Global consumption	4
TGBT1	113
TGBT2	120
PREDIS-MHI	91

Each notebook describes the area explored by it, with plans of the building, diagrams to illustrate its electrical system and tables that define to which load each meter is connected.



By choosing the year and the meter to explore by the user, the notebook calculates the average power and consumption of the chosen year. It also shows interactive load curves and monthly consumption in graphic form. The following figures illustrate, as examples, some images included in the notebook describing the TGBT1.

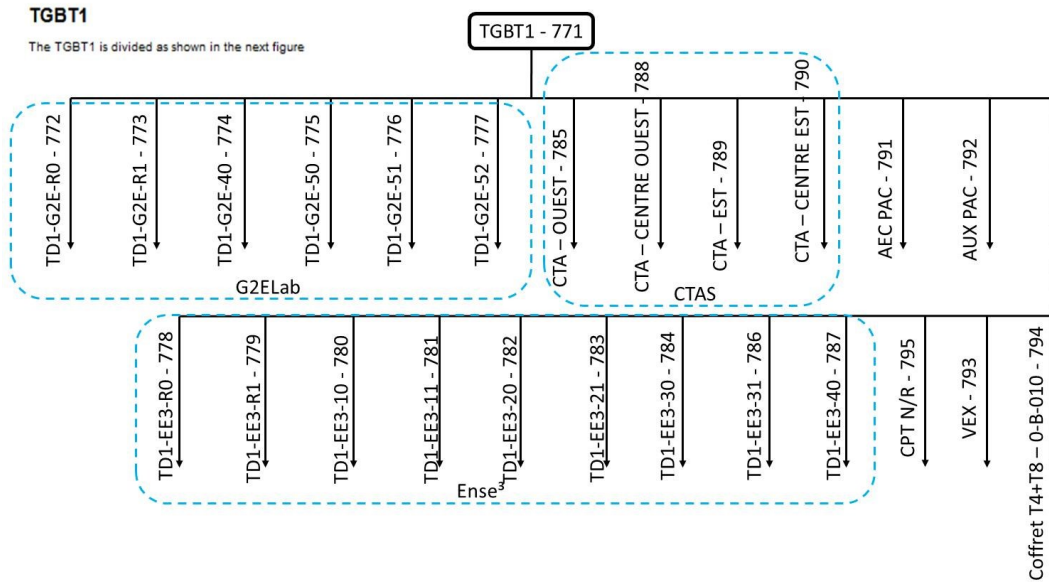


Figure 17 – Electric scheme of the TGBT1 present in it respective notebook.

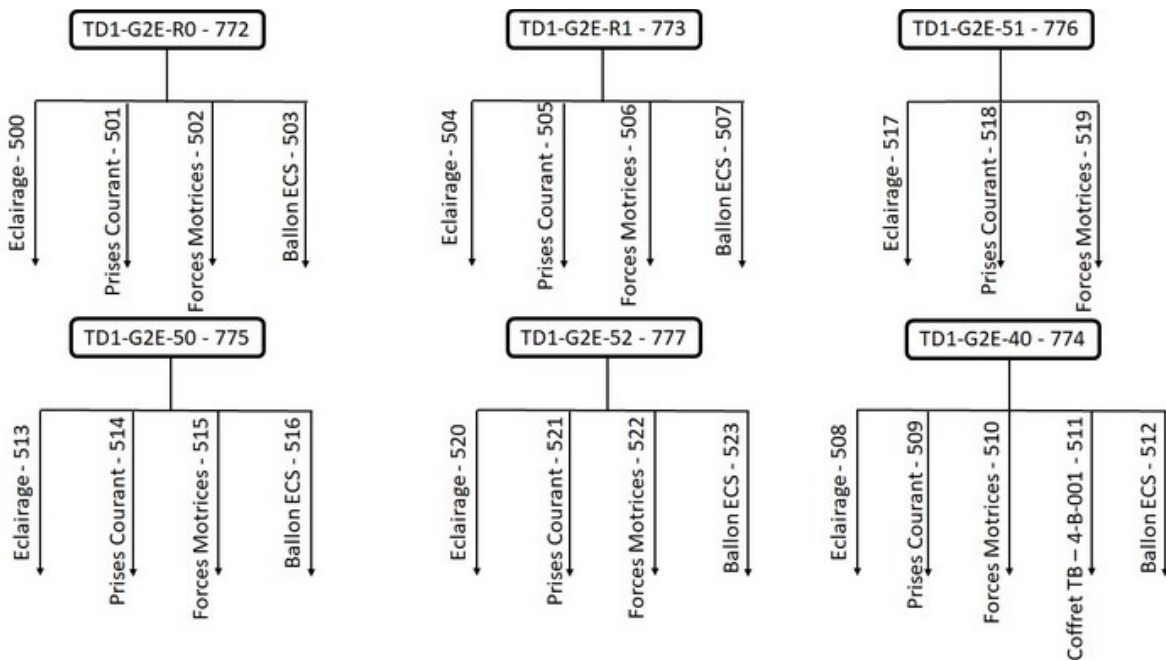


Figure 18 – Electric scheme of the switchboards connected to TGBT1 present in it respective notebook.

2nd floor

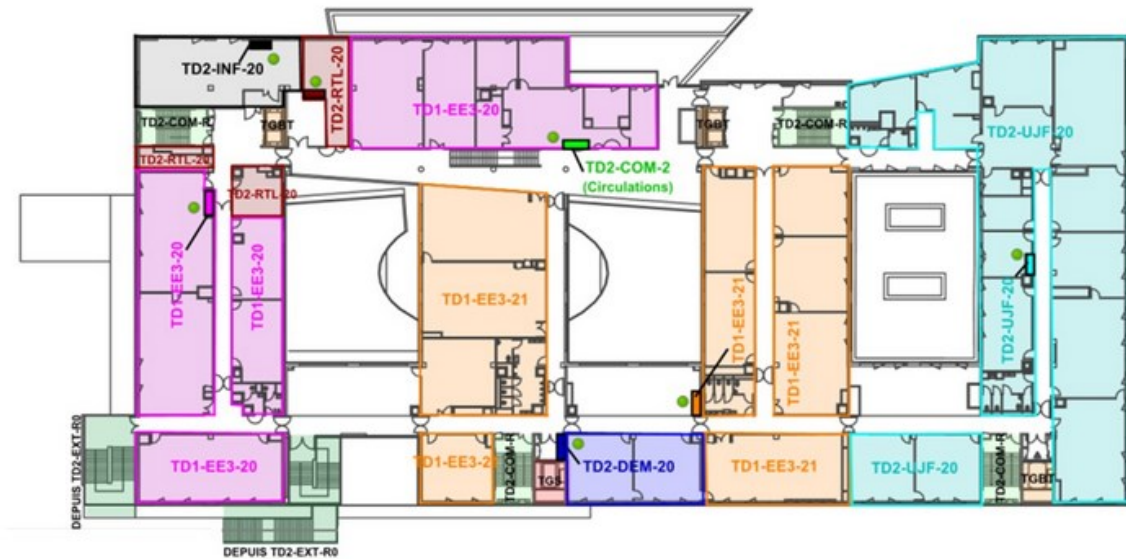


Figure 19 – Example of the floor plan of the 2<sup>nd</sup> floor of the GreEn-ER building with zones and their respective switchboards.

A couple of CSV files with the system design are also available. They are named "TGBT1\_n.csv", "TGBT2\_n.csv" and "PREDIS-MHI\_n.csv". In these files, each column stands for a switchboard. The head contains the names of the boards and the value in the first row represents the respective meter number. The values in the following rows represent the number of the sub-meters that are located downstream of the meter described in the first row. So, for example, in the file "TGBT1\_n.csv" there is a column that the head is "TD1-G2E-51". The value in the first row is "776", which represents the number of the meter of this switchboard. The values located in the following rows, "517", "518" and "519" represent the meters of the loads located downstream of the "TD1-G2E-51" switchboard.

## 2.4 GreEn-ER Electricity Consumption

The first step into the analysis of the energy consumption in a facility is to well understand its global consumption. In this way, it is possible to be aware of the order of magnitude of the consumption and visualize, for instance, seasonal patterns.

Figure 20 presents the monthly consumption of the GreEn-ER building during 2017. This year was chosen to be the main period of analysis because of its data quality, which will be addressed in Chapter 3 and 4.

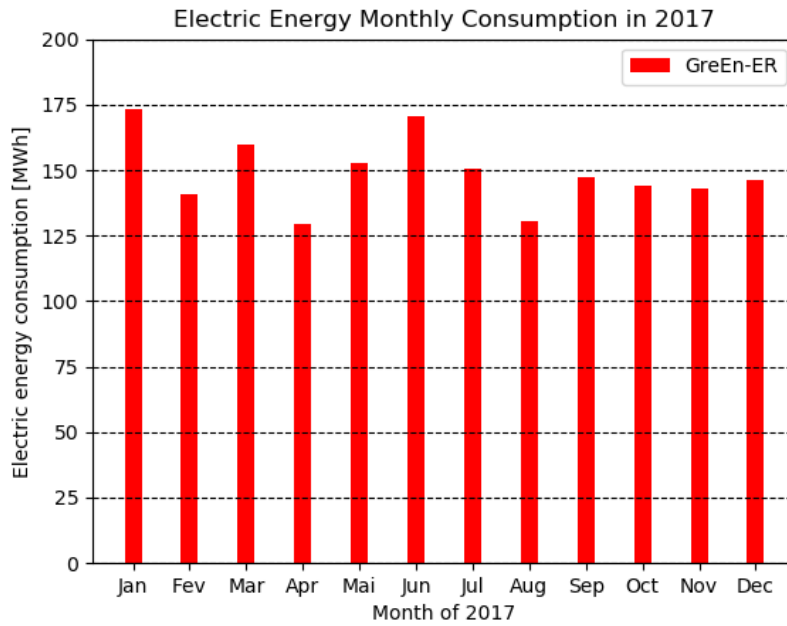


Figure 20 – Monthly electricity consumption of the GreEn-ER building in 2017.

The previous figure allows understanding the macro behavior of the consumption. It is possible to observe that the minimal consumption happens in April and August, periods with scholar vacations. In addition, the maximal consumption was in June, during the summer, when the cooling loads are more requested. In January, the consumption was also higher than other months, but this is due data quality problems that will be addressed in further chapters.

Another way to understand the electricity consumption is by the analysis of the load curve, which is presented in Figure 21, sampled at 1-minute time step.

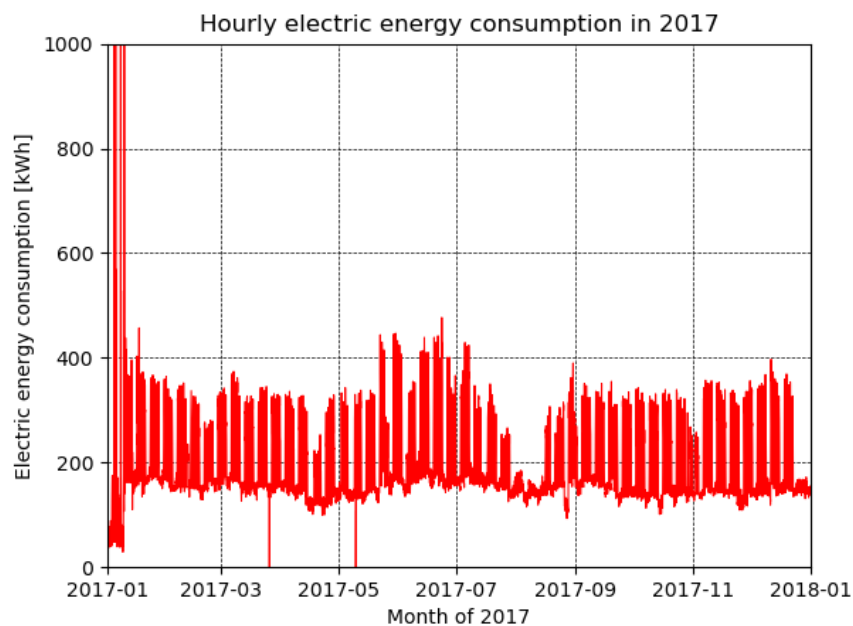


Figure 21 – Load curve of the GreEn-ER building in 2017.

It is possible to visualize, based on the previous figure, that there is a raise in the power consumption during the summer. Periods of scholar vacations, especially in August are also distinguishable. Furthermore, it is also possible to visualize some data quality problems, translated in the figure by the presence of outliers in the beginning of the series, in January. Those outliers may increase the monthly consumption, presented in Figure 20, and need to be treated carefully to more precise analysis. Additionally, it is possible to see that there is a weekly pattern, with the consumption dropping during weekends. Figure 22, presents a zoom in of Figure 21, in which it is possible to visualize the weekly and daily patterns. At the same time, Table 3 shows the average power according to the building’s occupancy, considering night time, weekends, holidays and vacations.

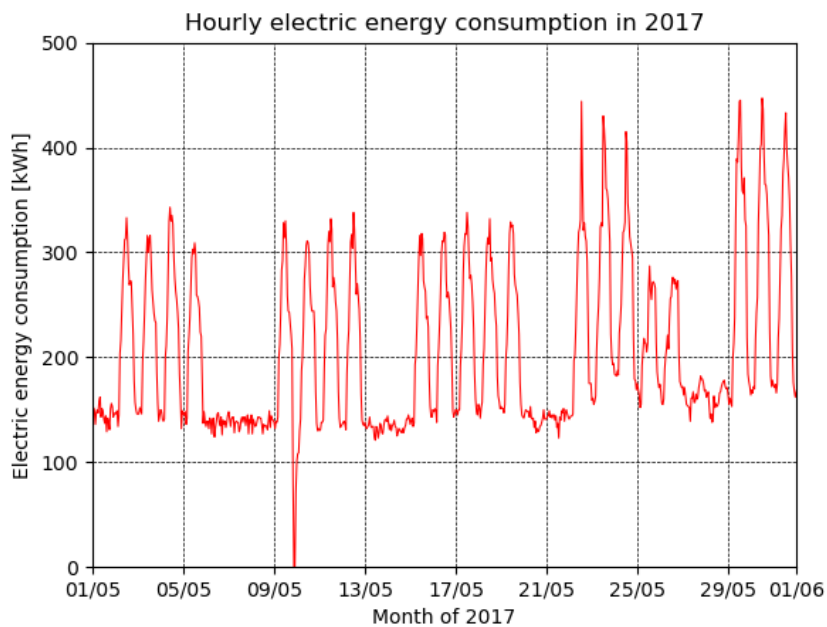


Figure 22 – Zoom in on the load curve of the GreEn-ER building in 2017.

Table 3 – Average power according to building’s occupancy.

Period	Average power [kW]
Whole year	204.11
Occupancy	287.46
Non-occupancy	158.89

The previously presented figure show the typical behavior of a tertiary sector building, with consumption concentrated during the daytime period on weekdays. However, it is also possible to visualize that the consumption when the building is unoccupied, at night and on weekends, corresponds to about 55% of the demand on the occupied periods. The high energy consumption during periods of non-occupancy evokes the need to take a closer look to the consumption exclusively during these periods, in order to provide a more detailed analysis of

the building's base load and thus facilitate the identification of potential opportunities to promote energy sobriety.

One way to understand how a facility, whether industrial or even a tertiary building, consumes energy is through a sectorial segregation of the consumption. In this way it is possible to identify the major consumers, from which the greatest savings potentials are usually extracted. In facilities where there was a massive deployment of remote meters, like in the GreEn-ER building, performing the consumption disaggregation is easier and it is strongly recommended to be done by all energy managers. However, this is not the rule for most facilities, in which consumption segregation is not trivial, and is not easily available.

Thus, the sectorial distribution of consumption was divided into three analyses. In the first, the disaggregation considering the whole period of operation is presented. The second analysis consists of the distribution of consumption taking into account only the periods of occupancy. Finally, the third analysis shows the sectorial distribution of consumption considering only the periods of non-occupancy of the building.

#### **2.4.1 Electricity Consumption Sectorial Distribution Considering the Whole Year**

As mentioned in section 2.3.1, the building's electric distribution is done by two main switchboards, called TGBT1 and TGBT2. Hence, the first step towards a sectorial distribution is to determine the consumption of each switchboards. According to Figure 23, TGBT1 is responsible for nearly 60% of the global consumption, while the remaining 40% is provided by TGBT2. The average power through the whole year is presented in Table 4.

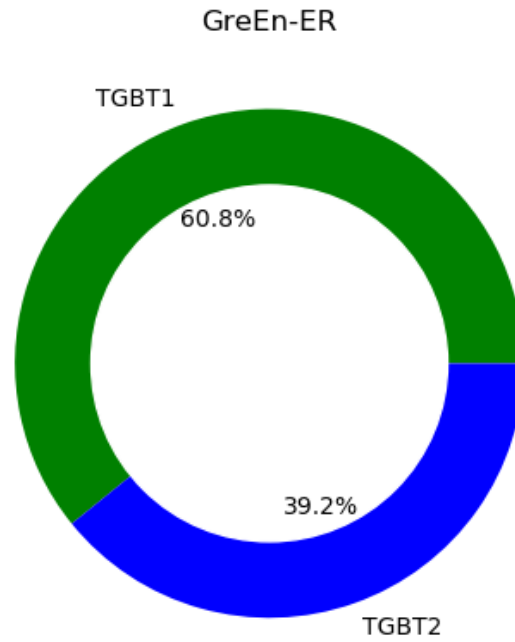


Figure 23 – Electricity distribution between TGBT1 and TGBT2.

Table 4 – Average power of TGBT1 and TGBT2.

Load	Average power [kW]	Share [%]
TGBT1	124.19	60.8
TGBT2	79.92	39.2
<b>GreEn-ER</b>	<b>204.11</b>	<b>100</b>

As it can be seen in Figure 11 and Figure 17, the main loads connected to TGBT1 are the switchboards that provide electricity to the G2Elab and to the Ense<sup>3</sup> areas, the heat pumps (AUX\_PAC and AEC\_PAC) and the Air Handling Units (AHUs). Additionally, the major consumer connected to this board is the datacenter (ARM\_NR). Figure 24 presents the contribution of each of these loads into the electricity consumption of TGBT1. The subsequent figures, on the other hand show the load curve of these loads, while their average power is presented in Table 5.

TGBT1

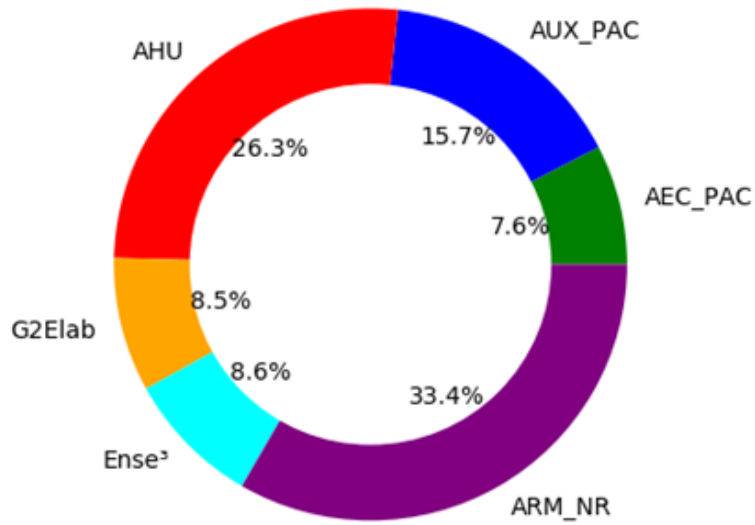
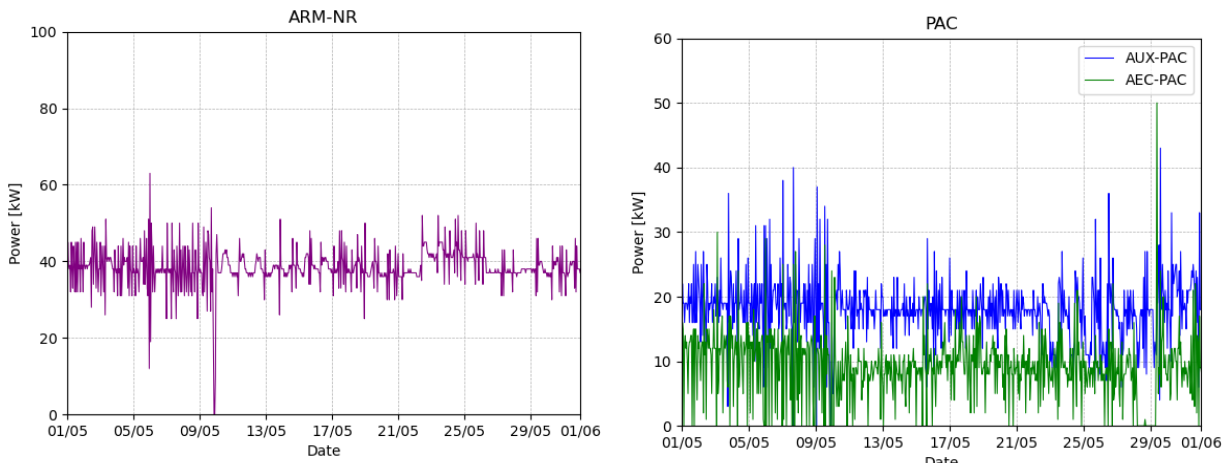


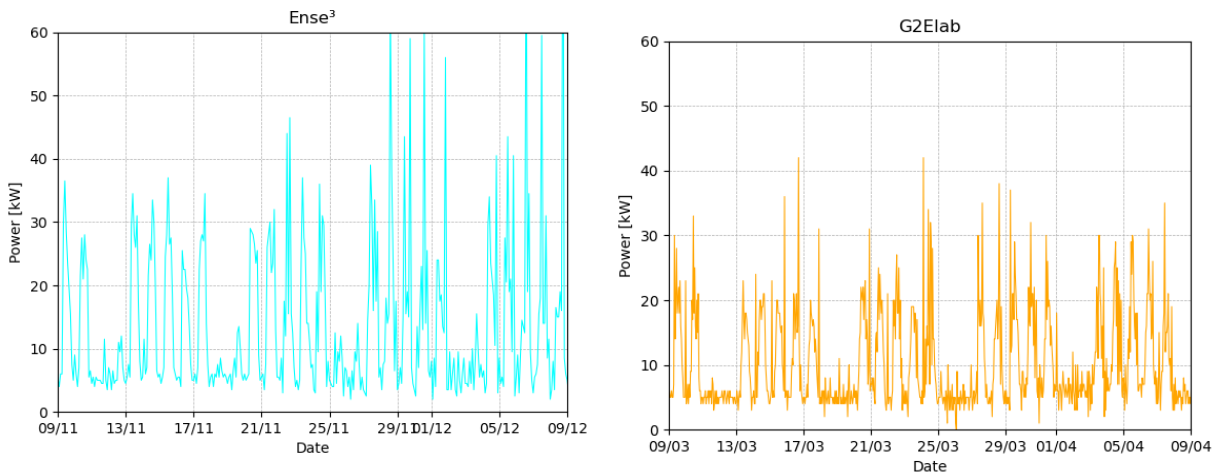
Figure 24 – Electricity distribution among main loads of TGBT1.



a) Datacenter (AMR\_NR)

b) ARM\_NR

Figure 25 – Zoom in on the load curve of the Datacenter and heat pumps system.



a) Ense<sup>3</sup>

b) G2Elab

Figure 26 – Zoom in on the load curve of the Ense<sup>3</sup> and G2Elab areas.

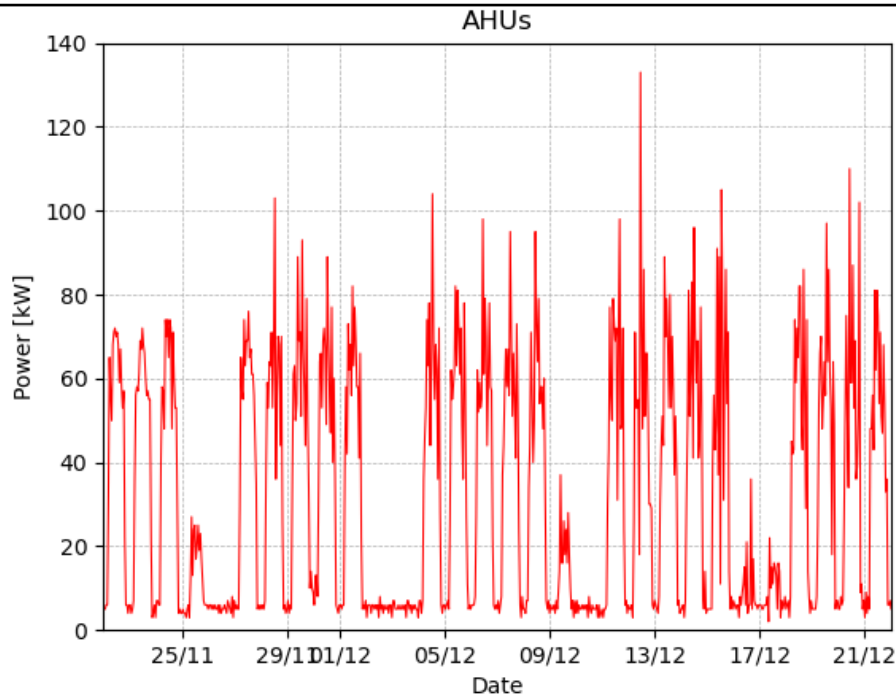


Figure 27 – Zoom in on the load curve of the AHUs.

Table 5 – Average power according of TGBT1 main loads.

Load	Average power [kW]
AHUs	31.58
ARM_NR	40.03
Ense <sup>3</sup>	10.30
G2Elab	10.17
AEC_PAC	9.10
AUX_PAC	18.85
<b>TGBT1</b>	<b>120.04</b>

By comparing Table 4 and Table 5, it can be seen that the values of average power corresponding to the TGBT1 do not match. This can be caused by some minor loads that are not monitored by the Building Management System, or by some error in the measurement system. Nevertheless, it corresponds to less than 4% of error and is judged acceptable.

Regarding TGBT2, Figure 28 presents the electricity consumption distribution among the main loads, in particular common areas, restaurant (Crous), cooling system (TD-GF), Predis area, air compressors and a special zone, that is more heavily monitored, called Predis-MHI. The subsequent figures, on the other hand, show the load curve zoomed in of these loads as examples, with their average power presented in Table 6.







Similarly to what happened to TGBT1, the values of average power of Table 4 and Table 6 for the TGBT2, do not match. However, the difference corresponds to less than 0.1% and is probably due to some error in the measurement system.

## 2.4.2 Electricity Consumption Sectorial Distribution according to occupancy periods

As mentioned earlier, the building's base load (consumption independent of the occupancy), corresponds to about 55% of the demand on the occupied periods. This share of the base load brings the need to also disaggregate the consumption by occupancy period. This may lead to the identification of energy sobriety opportunities, by pinpointing loads that do not follow the building's consumption pattern, without reducing their consumption during non-occupancy periods, for instance. Following tables present the average power of the main loads connected to both TGBT1 and TGBT2, according to occupancy periods. These tables may help identifying loads that do not reduce their consumption during non-occupancy periods.

Table 7 – Average power of TGBT1 according to occupancy periods.

Load	Average power [kW]		
	Total period	Occupancy period	Non-occupancy period
AHUs	31.58	59.48	16.44
ARM_NR	40.03	40.66	39.69
Ense <sup>3</sup>	10.30	16.66	6.85
G2Elab	10.17	15.34	7.37
AEC_PAC	9.10	9.64	8.79
AUX_PAC	18.85	19.12	18.70
<b>TGBT1</b>	<b>120.04</b>	<b>160.90</b>	<b>97.84</b>

Table 8 – Average power of TGBT2 according to occupancy periods.

Load	Average power [kW]		
	Total period	Occupancy period	Non-occupancy period
TD-GF	13.30	18.37	10.54
Crous	17.47	31.32	9.95
Predis	7.57	11.17	5.62
Common Areas	15.08	20.47	12.15
Predis-MHI	1.84	2.97	1.22
Others	5.69	7.62	4.63
Elevators	0.72	1.15	0.49
Air Compressors	18.21	18.10	18.27
<b>TGBT2</b>	<b>79.88</b>	<b>111.17</b>	<b>62.87</b>

From the previous tables, it is possible to identify some loads in which the reduction of average power, and consequently of electricity consumption, does not occur in periods of non-occupancy. This different pattern may happen because of misoperation of the load, or

misconception of the systems, or because their functioning does not depend on the building occupancy status. The latter may explain the pattern of the datacenter load (ARM-NR) that needs to continue its operation regardless the building occupancy. On the other hand, the fact that the compressed air system, and the heat pumps and ventilation (PAC and CVC) do not reduce their consumption during non-occupancy periods may represent misoperation of the loads or even misconception of the systems, unveiling energy efficiency and sobriety opportunities.

### **2.4.3 Energy Sobriety Opportunities**

As highlighted in previous sections, the analysis of the base load of the building may ease the way to the identification of energy sobriety opportunities. Major loads that do not follow the occupancy pattern may be operating in a non-optimized way, encouraging further analysis into the electricity consumption of these loads. One example of such loads is the compressed air system. As it can be seen in Figure 32, the energy consumption of the air compressors does not reduce during non-occupancy periods, such as weekdays nighttime periods or during weekends. The occurrence of this reduction during these periods, when the building occupancy is lower, could be expected for the compressed air system, since there would be fewer workstations using compressed air. This fact leads to closer investigation of this load. This section exposes in detail the electricity consumption of this load. First of all, the annual electricity consumption of the compressed air system was acquired in order to compare the year of 2017 to others. Figure 33 shows the annual electricity consumption of the compressed air system from 2017 to 2020.

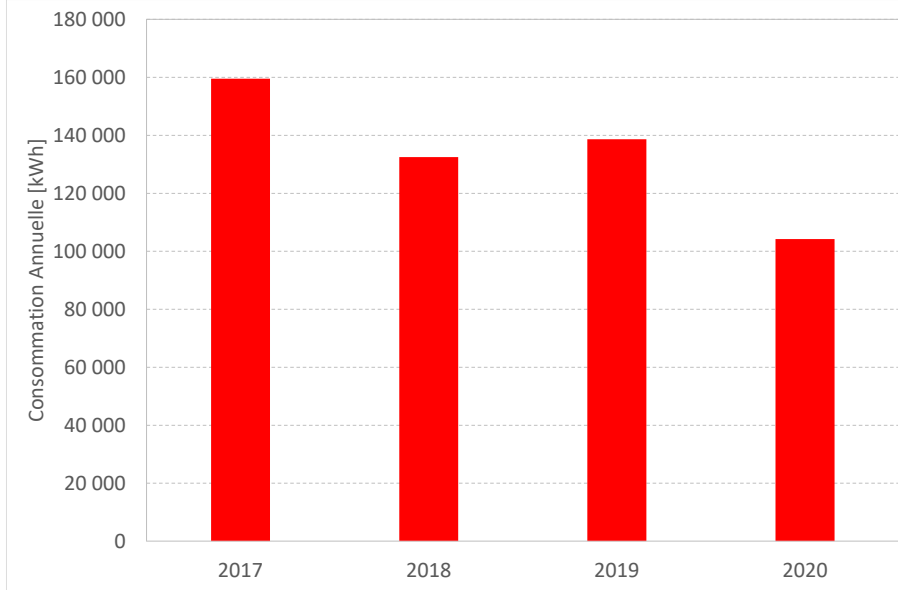


Figure 33 – Annual electricity consumption of the air compressors from 2017 to 2020.

The data exposed in the previous figure show that there was a reduction in the electricity consumption of this system in the later years. To understand if this reduction was due to compressed air demand decrease or if an energy efficiency measure was taken, it is necessary to break down the annual consumption into the monthly analysis, which can be seen in Figure 34. It is also important to highlight, that 2020 was an atypical year due to the COVID-19 lockdown ordered by French national authorities from mid-March to mid-May which could affect the consumption of the system.

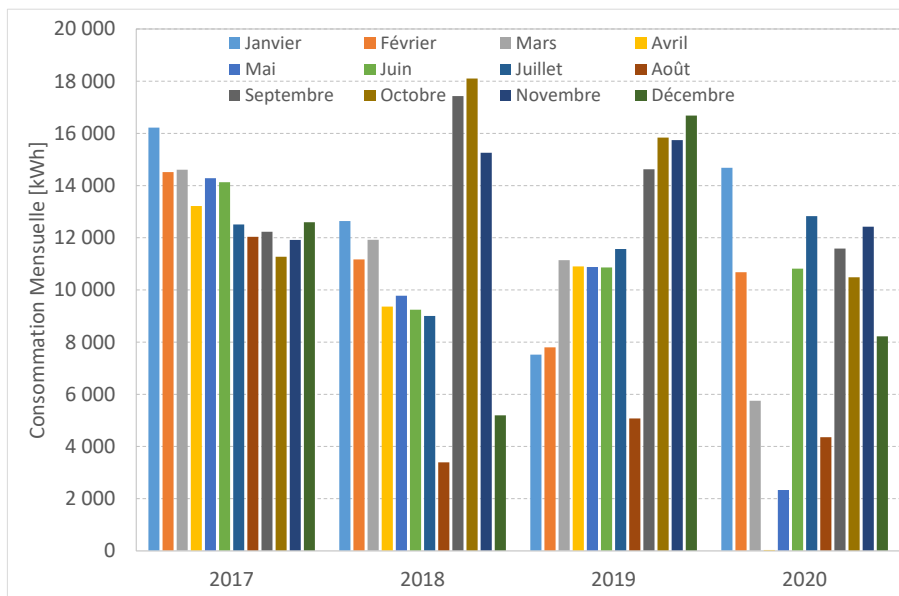


Figure 34 – Monthly electricity consumption of the air compressors from 2017 to 2020.

As it can be observed from Figure 34, some energy sobriety measures were taken during vacation periods, mostly in August (a period of non-occupancy), from 2018 onward, when the consumption was notably smaller than the other months. Also, it can be seen, in 2020, especially during April, that the system was apparently shutdown during the COVID-19 lockdown. This fact indicates that it is possible to reduce the compressed air system demand, and thus the electricity consumption, during non-occupancy periods without undermine the consumers. Hence, the next step is to visualize if this reduction already happened from 2018 onward. This can be analyzed by breaking down the monthly consumption into a weekday analysis that can be seen in Figure 35.

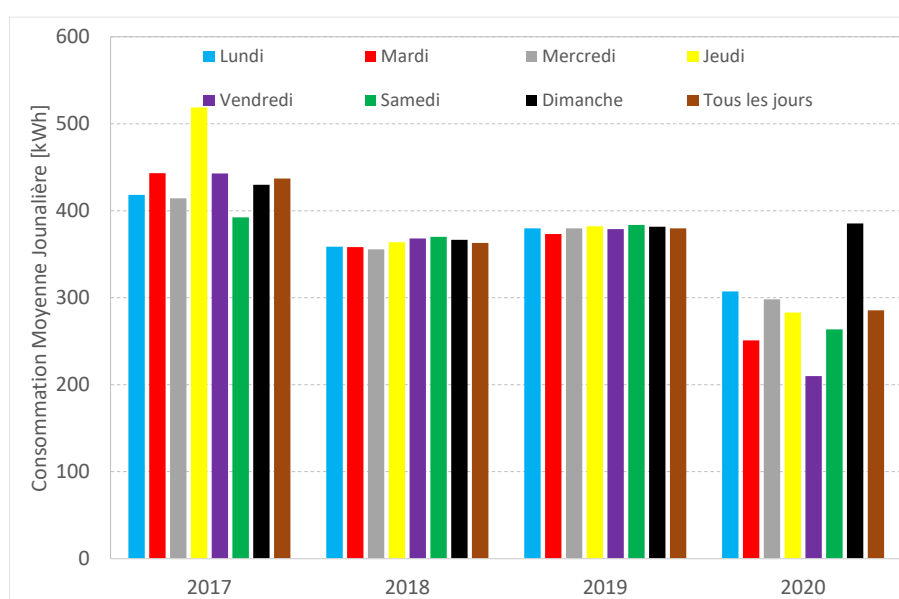


Figure 35 – Average electricity consumption of each weekday of the air compressors from 2017 to 2020.

As it can be observed in Figure 35, there is no difference between weekdays and weekends regarding the consumption of the compressed air system. In 2017, the consumption in the Saturdays was similar to the Wednesday one, while the Sunday consumption was nearly the same as at the Fridays. During 2018 and 2019, the consumption was nearly the same for all the days of the week, while during 2020, the higher consumption happened during Sundays, while the lower one was on Fridays, although it is important to remember the atypical situation of 2020 due to COVID-19 restraints. To illustrate what happened during the vacation period in 2018, Figure 36 shows the load curve of the compressed air system during August of 2018.

As it can be seen in Figure 36, during the vacation period of 2018, from August 1<sup>st</sup> to August 22<sup>nd</sup>, the compressed air system was almost shutdown. Its average power during this

period was 0.45 kW. These values indicate that it is possible to almost shutdown the compressed air system during non-occupancy periods of the building without undermine the consumers.

To estimate the savings obtained by shutting down the compressed air system during non-occupancy periods, it is enough to consider that the average power of the system during non-occupancy periods is equal to 0.45 kW, which was the average value during the summer vacation in 2018. Considering the nighttime of weekdays, weekends, holidays and vacations, the occupancy period lasts 3081 hours per year, while during the 5679 remaining hours, the building is not occupied. Table 9 presents the annual electricity consumption of the compressed air system if the average power of the air compressors was the same as the power during the vacation time of 2018. At the same time, Table 10 shows the savings that would have been achieved for 2017, 2018 and 2019 in the same situation.

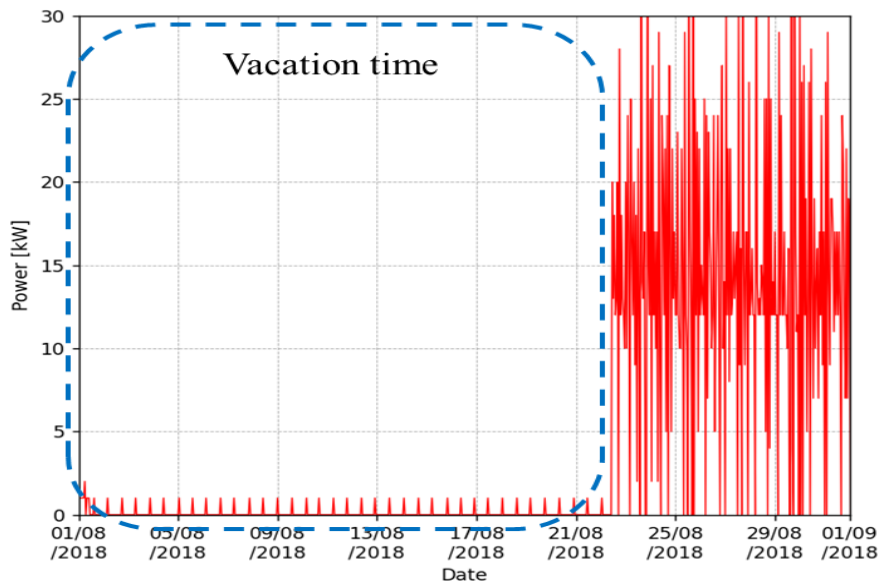


Figure 36 – Compressed air system load curve during vacation time in 2018.

Table 9 – Annual compressed air system electricity consumption considering the shutdown of the system during non-occupancy periods.

	Occupancy period	Non-occupancy period	Total period
Average power [kW]	18.10	0.45	6.66
Hours	3081	5679	8760
<b>Annual electricity consumption [MWh]</b>	55.77	2.56	58.32

Table 10 – Annual compressed air system electricity consumption considering the shutdown of the system during non-occupancy periods.

Year	Annual electricity consumption [MWh]		Savings	
	Before system modification	After system modification	[MWh]	[%]
2017	159.52	58.32	101.20	63%
2018	132.49	58.32	74.17	56%
2019	138.63	58.32	80.31	58%

The data presented in the previous tables show that it could be possible to achieve nearly 60% of savings in the compressed air system if the compressors' demand during all periods of non-occupancy was equal to the one during the summer vacations of 2018. Even if the savings in this chapter are overestimated, the analysis itself lead to further investigations, in which more savings opportunities may be found.

## 2.5 Conclusions

Data-driven approaches combined with machine learning techniques are more and more frequent in the energy sector. Because of that, the importance of the availability of datasets containing real measurement data of energy consumption is in constant increase. Nevertheless, most of datasets available nowadays concern residential environments, which have specific loads and functioning patterns, when compared to the tertiary sector. The availability of those datasets moved forward the research in the residential sector in some areas such as the Non-intrusive load monitoring, leaving the tertiary sector behind. Thus, one objective of this thesis was to make a dataset available in open-data, described in this chapter, containing the electricity consumption of the GreEn-ER building, both aggregated and disaggregated in several levels. Because of the completeness of the data, which will be addressed in the next chapter, the data available in the dataset concerns mostly 2017.

By using the data accessible from the GreEn-ER dataset it was possible to analyze the buildings consumption in more detail. For instance, the base load of the building (consumption during non-occupancy periods, such as at night and on weekends) corresponds to about 55% of the demand on the occupied periods. This high share brings attention to this specific period and potential energy sobriety opportunities. One way to try to identify some potential savings is to disaggregate the electricity consumption and identify the major consumers of the building. Some of them are the compressed air system, the AHUs, the restaurant (Crous) and the Datacenter (ARM\_NR).

Among these loads, there are a few that do not change their functioning mode during non-occupancy periods as it was expected. That is the case of the air compressors. During periods when fewer people are inside the building, the consumption of compressed air, and hence the electricity consumption of the air compressors, is expected to decrease. However, that is not what happens during 2017. Furthermore, during 2018 and 2019, the system was almost shutdown during the summer vacations. This fact may indicate that it is possible to decrease the compressed air consumption during non-occupancy periods without undermine



the workstations that use this resource. A previous analysis suggests that savings of nearly 60% are achievable in the compressed air system by decreasing its consumption, during non-occupancy periods.

## References

- [1] Hart G. W., "Nonintrusive appliance load monitoring," Proc. IEEE, vol. 80, no. 12, Dec. 1992 pp. 1870–1891.
- [2] Kolter, J.Z.; Johnson, M.J. 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.
- [3] Anderson, K., Ocleanu, A., Benitez, D., Carlson, D., Rowe, A., Berges, M. 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), 2012. 1-5.
- [4] Zimmermann, J. P., Evans, M., Griggs, J., King, N., Harding, L., Roberts, P., Evans, C. Household electricity survey a study of domestic electrical product usage. Household Electricity Survey: A Study of Domestic Electrical Product Usage, 2012.
- [5] Reinhardt, A., Baumann, P., Burgstahler, D., Hollick, M., Chonov, H., Werner, M., Steinmetz, R. On the accuracy of appliance identification based on distributed load metering data. 2012 Sustainable Internet and ICT for Sustainability, SustainIT 2012. 2012.
- [6] Makonin, S., Popowich, F., Bartram, L., Gill, B. , Bajic, I. V. AMPds: A public dataset for load disaggregation and eco-feedback research. Electrical Power and Energy Conference (EPEC), 2013 IEEE, 1-6. 2013.
- [7] Batra, N., Gulati, M., Singh, A., Srivastava, M. B. It's Different: Insights into home energy consumption in India. Proceedings of the Fifth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (ACM BuildSys). 2013.
- [8] Maasoumy, M., Sanandaji, B., Poolla, K., Vincentelli, A. S. Berds-berkeley energy disaggregation dataset. Proceedings of the Workshop on Big Learning at the Conference on Neural Information Processing Systems (NIPS), 2013.
- [9] Gisler, C., Ridi, A., Zufferey, D., Khaled, O. A. , Hennebert, J. Appliance consumption signature database and recognition test protocols. 2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), 12-15 May 2013. 336-341.
- [10] Kelly, J. , Knottenbelt, W. The UK-DALE dataset, domestic appliance-level electricity demand and whole house demand from five UK homes. Scientific Data, 2. 2015.
- [11] Beckel, C., Kleiminger, W., Cichetti, R., Staake, T. , Santini, S. The ECO data set and the performance of non-intrusive load monitoring algorithms. BuildSys 2014 - Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, 2014a. 80-89.
- [12] Monacchi, A., Egarter, D., Elmenreich, W., D'alessandro, S., Tonello, A. M.. GREEND: An energy consumption dataset of households in Italy and Austria. The 5th IEEE International Conference on Smart Grid Communications (SmartGridComm). 2014
- [13] Pereira, L., Quintal, F., Gonçalves, R. , Nunes, N. J. SustData: A public dataset for ICT4S electric energy research. ICT for Sustainability 2014, ICT4S 2014, 2014. 359-368.
- [14] Batra, N., Parson, O., Berges, M., Singh, A., Rogers, A. 2014b. A comparison of non-intrusive load monitoring methods for commercial and residential buildings. arXiv preprint arXiv:1408.6595.
- [15] Gao, J., Giri, S., Kara, E. C., Berg, M., #233 2014. PLAID: a public dataset of high-resolution electrical appliance measurements for load identification research: demo abstract. Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings. Memphis, Tennessee: ACM.
- [16] Akshay, U., Nambi, S. N., Lua, A. R., Prasad, R. V. Loced: Location-aware energy disaggregation framework. Proceedings of the Second ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (ACM BuildSys). 2015.
- [17] Parson, O., Fisher, G., Hersey, A., Batra, N., Kelly, J., Singh, A., Knottenbelt, W., Rogers, A.

- Dataport and NILMTK: A building data set designed for non-intrusive load monitoring. 2015 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2015, 2015. 210-214,
- [18] Picon, T., Meziane, M. N., Ravier, P., Lamarque, G., Novello, C., Bunetel J.-C. L., Raingeaud, Y.. COOLL: Controlled On/Off Loads Library, a Public Dataset of High-Sampled Electrical Signals for Appliance Identification. 2016. arXiv preprint arXiv:1611.05803.
- [19] Kahl, M., Haq, A. U., Kriechbaumer, T., Jacobsen, H.-A. Whited-a worldwide household and industry transient energy data set. Workshop on Non-Intrusive Load Monitoring (NILM), 2016 Proceedings of the 3rd International, 2016.
- [20] Murray, D., Stankovic, L., Stankovic, V.. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. *Scientific Data*, 4, 160122, 2017.
- [21] Martin Nascimento, G. F., Delinchant, B., Wurtz, F., Kuo-Peng P.; Jhoe Batistela, N. GreEn-ER Data - Jeux de données de consommation d'électricité dans le secteur tertiaire. 2021. <hal-03322581>
- [22] Delinchant, B.; Wurtz, F.; Ploix, S.; Schanen, J.; Marechal, Y. GreEn-ER living lab: A green building with energy aware occupants. in 2016 5th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS), Rome, 2016, pp. 1-8.
- [23] Martin Nascimento, G. F.; Delinchant, B.; Wurtz, F.; Kuo-Peng, P.; Jhoe Batistela, N.; Laranjeira, T. GreEn-ER - Electricity Consumption Data of a Tertiary Building. *Mendeley Data*, V1, 2020 <http://dx.doi.org/10.17632/h8mnmthn5w.1>
- [24] Seagate, Reinsel, D., Gantz, J., Rydning, J. The Digitization of the World From Edge to Core. <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-data-age-whitepaper.pdf>, 2018
- [25] Ministère de l'Enseignement Supérieur, de la Recherche et de l'innovation, France. Plan national pour la science ouverte, 2018.
- [26] Austin, C.C., Bloom, T., Dallmeier-Tiessen, S. et al. Key components of data publishing: using current best practices to develop a reference model for data publishing. *Int J Digit Libr* 18, 2017, 77–92. <https://doi.org/10.1007/s00799-016-0178-2>

### 3 Data Quality in Energy Domain – The GreEn-ER Case

Unreliable data are as harmful as not having data at all. The increase of the use data-driven approaches raises a concern about the quality of the data used. Therefore, this chapter presents a brief overview about data quality, in which data quality problems and dimensions are discussed, especially in the energy domain. In addition, this chapter also presents an assessment on the data quality of the GreEn-ER dataset, especially in terms of accuracy and completeness. Besides that, an example of the issues that poor data quality can cause in supervised machine learning techniques is presented.

#### Chapter Contents

3.1 Overview of Data Quality Problems .....	68
3.2 Data Quality Assessment of the GreEn-ER Dataset .....	69
3.3 Poor Data Quality Influence in Machine Learning .....	76
3.4 Conclusions.....	79

### 3.1 Overview of Data Quality Problems

Currently, in practically all domains, decisions are made based on data. From policies making to daily-bases decisions, such as which route to take to commute or city zoning plan, people rely on data. However, making decisions relying on unreliable data can be more catastrophic than blind decisions. Because of that, it is important to be able to evaluate the quality of the data used to make a decision.

Data quality (DQ) assessments can be differentiated in objective and subjective [1]. An objective DQ evaluation rely on database integrity rules, which can be used by software systems to measure the quality of datasets. In contrast, a subjective data quality assessment is based on user feedbacks and reflect the needs and experiences of stakeholders. There are typically conducted through surveys or interviews to evaluate the quality of data products from the consumers’ perspective [2].

Data quality problems can take many forms, and each type of problem demands a different solution for its mitigation [3]. Because of these many forms that DQ problems may assume, it is necessary to well classify them in order to know how to tackle these problems. The categories into which data quality problems have been classified are called Data Quality dimensions [4] [5] [6]. Table 11 presents the definition of 17 DQ dimensions defined in Ge’s work [7].

Table 11 – Data quality dimensions. [7]

<b>DQ Dimension</b>	<b>Description</b>
Accessibility	Accessible, obtainable, retrievable, available.
Security	Secure, protected, authorized access.
Relevancy	Useful, relevant, applicable, helpful.
Value-added	Beneficial, valuable, add value to operations.
Accuracy	Correct, accurate, free of error, precise.
Completeness	Sufficient, complete, comprehensive, include all necessary values, detailed.
Timeliness	Current, up to date, delivered on time, timely.
Consistency	Consistent meaning, consistent structure, presented in the same format.
Interpretability	Interpretable, without inappropriate language and symbol, readable.
Objectivity	Impartial, unbiased, objective, based on facts.
Representation	Concise, compact.
Reliability	Reliable, dependable.
Believability	Believable, trustworthy, credible.
Reputation	From good sources, of good reputation, well referenced.
Ease of Manipulation	Easy to manipulate, easy to aggregate, easy to combine.
Ease of Understanding	Easy to understand, easy to comprehend, easy to identify the key point.
Appropriate Amount	Not too much, not overload, not too little.

However, DQ problems are usually field-specific, i.e. in each domain there are DQ

dimensions that are more critical, because of the type of data generated. Considering the specificity of the energy domain, seven data quality dimensions, listed as follows, become more important [8] [9].

- Accessibility
- Accuracy
- Completeness
- Consistency
- Timeliness
- Interpretability
- Believability

The accessibility dimension deals with how easy data are retrieved. In the context of an open-science environment, the accessibility increases its importance. The accuracy dimension quantifies if the measured data are correct. For instance, outlier data can be classified into the accuracy dimension. Completeness deals mostly with missing data. Missing samples can be often due to device malfunctioning, or even communication problems between the meters and the data storage application.

At the same time, data consistency is also important in the energy domain. It is usual to get data from different devices, which need to output data in a compatible format to allow analyses using the measured data. In addition, in the energy domain, several analyses often deal with time series data. Hence, issues with the timestamp records may complicate these analyses. The dimension that assesses if the data are correctly sampled is the timeliness.

Moreover, because of its impact to economic benefits in the case of energy efficiency analyses, trustful data are vital, making believability another important dimension in the energy domain. Likewise, interpretability is another important dimension. Measurement data often present corrupted data that prevents the use of these data.

### **3.2 Data Quality Assessment of the GreEn-ER Dataset**

It is unusual that a dataset does not present data quality problems in some instance, and the GreEn-ER dataset is no exception. Some examples were already exposed in Chapter 2, in the form of the presence of outliers, which configures an accuracy dimension problem. Another problem that arises from the GreEn-ER dataset is the lack of completeness, which sometimes is also the cause of the presence of outliers. The problem of incomplete data tends

to arise when the communication between the meters and the data storage system is temporary lost. To address these eventual problems, the measuring and storing data system were designed to evaluate the energy consumption rather than the load curve. That way, in a determined period, the final energy consumption tends to be accurate, because the energy meter is capable to integrate the measured value and send the accumulated value once the communication is restored. Sometimes it represents a high peak of consumption in a short time step. This behavior is not critical when dealing only with the energy consumption, but it becomes problematic when there is an interest in reconstructing the power pattern of the loads.

As an example, Figure 37 presents 15 days' worth of data from the TGBT1, showing both energy consumption and power (retrieved from energy). It can be seen in this figure two of the main data quality problems present in the GreEn-ER dataset, lack of completeness and poor accuracy, represented by the peaks.

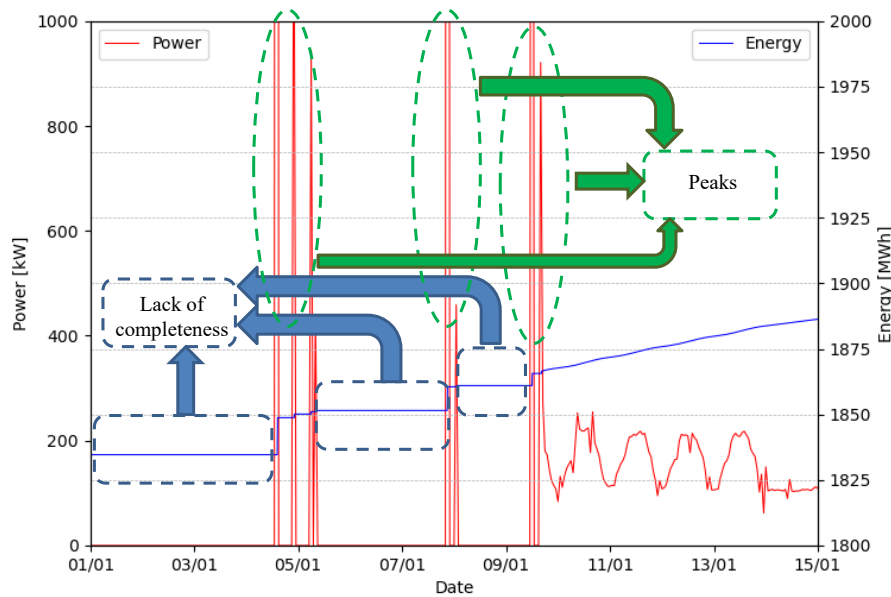


Figure 37 – Zoom in of the TGBT1 load curve highlighting some data quality problems.

### 3.2.1 Accuracy

Figure 37 have presented two main data quality problems. One of them corresponds to data accuracy, expressed through the presence of outliers. Nevertheless, accuracy problems can appear in more than one form. In an energy dataset, with several layers of aggregated and disaggregated data, it is important to check if the sum of the consumption measured by all downstream meters is compatible with consumption measured by the upstream meter. Table 12 presents some examples of this comparison. In this table, real meter represents the upstream meter while virtual meter represents the sum of all downstream meters in the layer immediately below.

Table 12 – Example of comparison between real and virtual meters of GreEn-ER.

Switchboard	Annual energy consumption [MWh]			Error [%]
	Real Meter	Virtual Meter	Error	
TGBT1	1087.93	1053.74	34.20	3.14
TGBT2	700.10	699.66	0.44	0.06
TD-GF	116.48	115.58	0.90	0.77
TD1-G2E-52	31.63	29.61	2.03	6.40
TD1-EE3-10	17.08	16.69	0.39	2.31
ARM_NR	350.64	345.15	5.48	1.56
TD2_COM_S	46.16	45.68	0.48	1.04
TD2_PRE_40	33.52	32.70	0.83	2.46
TD2_DEM_40	16.05	14.83	1.23	7.64

The data presented in the previous table support the choice of logging measurements of consumed energy rather than electrical power. The energy meters installed in the building have the ability of integrate the consumed energy value. Therefore, even during an eventual communication failure, the value of consumed energy would be correct when the communication is restored.

Nevertheless, the power, expressed in the form of load curves allows further analyses. It is easier to detect consumption patterns, outliers and even lack of completeness by analyzing load curves rather than cumulative energy graphics. In addition, non-intrusive load monitoring techniques, used in further chapters of this work, rely on load curves to disaggregate the overall consumption down to the appliance level. Therefore, it is also important to assess the accuracy of the GreEn-ER data in terms of power. Because of the 1kWh resolution of the energy meters, small loads are not the most suited for these analyses.

Hence, the data from the TGBT2, and the first layer of meters immediately downstream to them were taken as an example. The idea is to assess the accuracy of a virtual meter, formed by the sum of the loads of the first layer, downstream the TGBT2. This virtual meter is then compared with the real meter corresponding to the TGBT2 one, which is



assumed to be the ground truth. Figure 38 shows the comparison between the real and virtual meters of the TGBT2, using 1 hour as sampling interval, while Table 13 exposes the average power over one year.

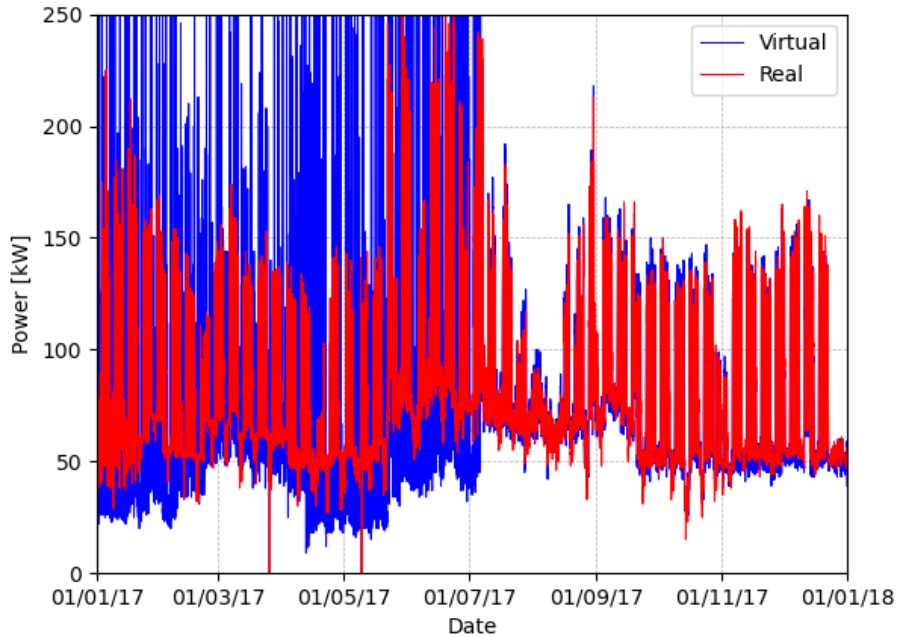


Figure 38 – Real and virtual meters of TGBT2 in terms of power.

Table 13 – Comparison between real and virtual meters considering the average power over one year.

Switchboard	Average power [kW]			Error [%]
	Real Meter	Virtual Meter	Error	
TGBT2	79.92	79.87	0.05	0.06

Although the values presented in Table 13 show that the error of the average power over one year between the real and the virtual meters of the TGBT2 is practically irrelevant, less than 0.1%, the load curves presented in Figure 38 show that the samples values do not match, especially before July.

One way to measure the accuracy of the virtual meter is to count the number of samples whose percentage error, calculated by the following equation, from the real meter is greater than a certain tolerance.

$$Percentage\ Error\ [\%] = \frac{|Real - Virtual|}{Real} * 100 \quad \text{Equation 1}$$

It can be inferred that this accuracy measure is highly dependent on the sampling interval, since the energy consumed at the end of the series is quite similar. Because of that, the accuracy was calculated considering seven sampling intervals: 10 minutes, 15 minutes, 30

minutes, 1 hour, 6 hours, 12 hours and 1 day. The tolerance varies from 10% up to 100%. The results are presented in Table 13 and illustrated in Figure 39.

Table 14 – TGBT2's virtual meter accuracy as a function of the sampling interval and the tolerance allowed.

Tolerance [%]	Accuracy [%]						
	Sampling interval						
	10 minutes	20 minutes	30 minutes	1 hour	6 hours	12 hours	1 day
10	30.12	34.76	44.58	49.28	56.25	63.57	70.30
20	52.63	56.75	62.97	64.29	68.67	76.67	81.74
30	66.51	69.33	73.45	74.71	80.55	84.99	90.46
40	75.92	78.40	82.46	83.16	86.83	91.54	95.64
50	83.53	86.20	90.54	91.25	93.52	94.68	97.55
60	87.78	90.67	94.90	95.99	96.11	95.63	98.64
70	91.58	93.89	96.99	96.72	96.52	96.73	98.64
80	94.23	95.86	97.54	96.98	96.72	97.27	99.18
90	95.66	96.72	97.68	97.21	96.93	97.54	99.46
100	96.48	97.16	97.80	97.38	97.27	98.36	99.46

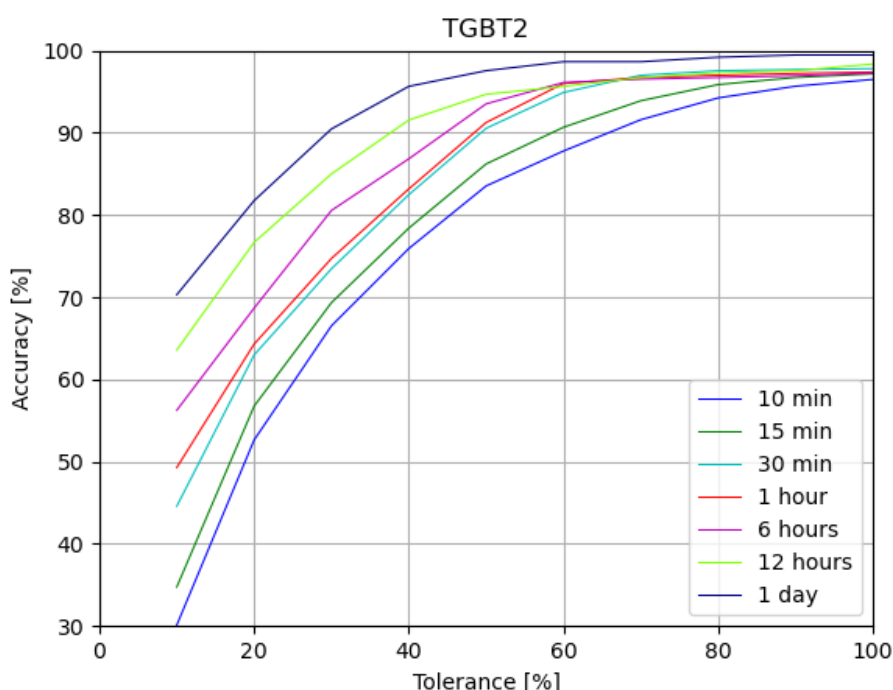


Figure 39 – TGBT2's virtual meter accuracy as a function of the sampling interval and the tolerance allowed

The results presented in the previous table and figure show that, in the case of the GreEn-ER dataset, the smaller sampling interval is not always the best choice, especially when DQ issues are present. Bigger sampling intervals, using the mean as resampling function, smooth the curves, decreasing the peaks that are clearly DQ issues and increasing the reliability of the data.

### 3.2.2 Completeness

Another data quality problem present in the GreEn-ER dataset is the lack of completeness. This problem is usually expressed by missing samples generating gaps in a time series, but it can appear in more than one way.

In the GreEn-ER data set, lack of completeness is expressed in two different ways. One is the presence of null samples. The presence of such samples forms a gap in the time series. This problem was identified in a chiller load, the *Groupe Froid 1*, responsible for the cooling of the building in summer. This meter is located in the second layer downstream the TGBT2 one. Figure 40 presents the location of the meters related to this load, while Figure 41 presents its load curve highlighting the completeness problem.

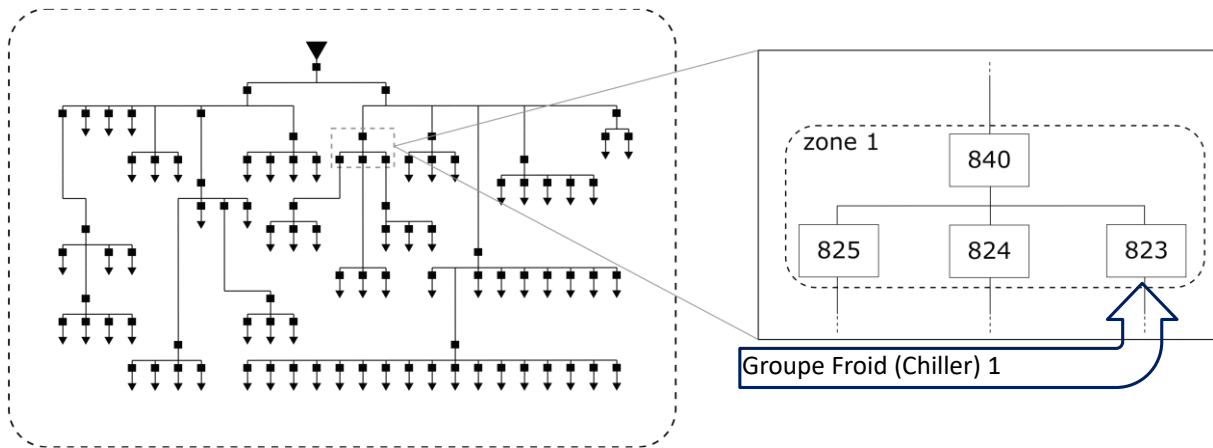


Figure 40 – Location of the Groupe Froid 1 meter.

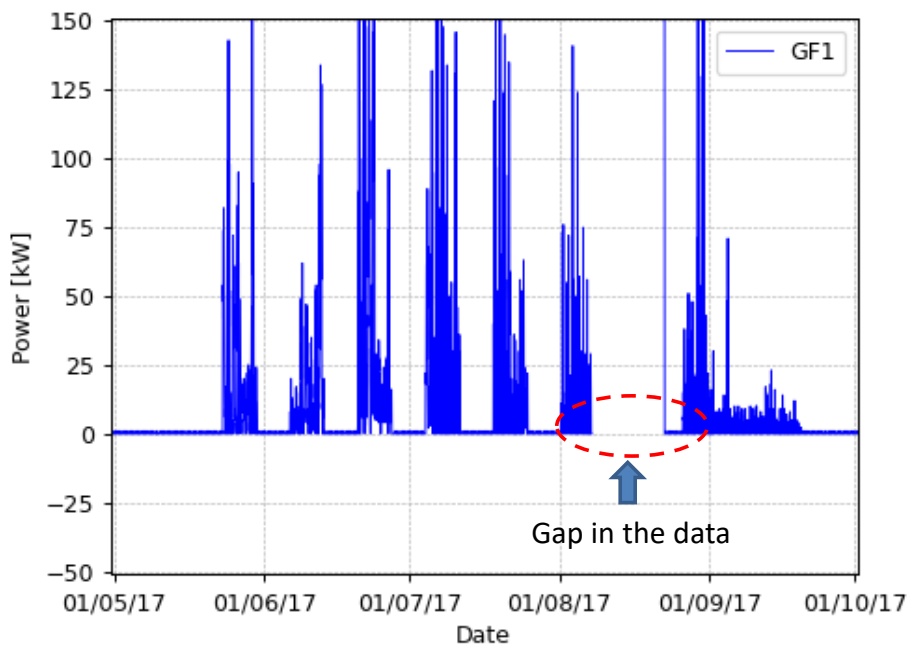


Figure 41 – Load curve of the Groupe Froid 1 highlighting a gap in the data.

The total completeness of the GreEn-ER dataset can be defined as the percentage of missing samples contained within the dataset. Hence, considering 10 minutes sampling interval and 2017 as base year, all missing samples were counted. The results are presented in the following table.

Table 15 – Completeness due missing samples.

Switchboard	Number of meters	Total number of samples	Missing Samples	Completeness [%]
TGBT1	114	6 043 595	61 388	98.98
TGBT2	122	6 464 019	235 303	96.36
PREDIS-MHI	91	4 782 323	0	100.00
<b>Total</b>	<b>327</b>	<b>17 289 937</b>	<b>296 691</b>	<b>98.28</b>

The results presented in Table 15 show that almost 300 000 samples are missing in the GreEn-ER dataset, which represents 1.72% of all data. Numerically, it may not represent a big loss of data; however, when there are several consecutive missing samples, there is a big loss of information, as shown in Figure 41. In addition, among the missing samples there are meters without any data. Hence, an application to monitor data quality, indicating to a supervisor when there is a faulty device, would be an enhancement to the Building Management System, reducing loss of information and maintenance costs.

Nevertheless, in the case of the GreEn-ER dataset, the completeness problem is expressed in one more way. As mentioned before, when the communication between the meters and the data storage system is temporary lost, the energy meter is able to integrate the measured value of consumed energy and send the accumulated value once the communication is restored. In such cases, the storage system keeps recording the last value received in the subsequent samples. This leads to the steps, in terms of energy, and peaks, in terms of power, seen in Figure 37. However, this problem is harder to quantify. Because of the 1kWh resolution of the energy meters, it is not trivial to discern the zero consumption due to the communication issue or zeros due to the meters latency, especially in the case of small loads, in which the accumulated consumption logged after the communication issue is not that different from the normal consumption.

Hence, an algorithm to identify these peaks, also known as outliers, even the local ones, for the main loads of the GreEn-ER building was developed and is detailed in the next chapter.

### 3.3 Poor Data Quality Influence in Machine Learning

It is not a secret that machine-learning techniques, especially the supervised ones, rely on data to their training. Hence, it is only logical that poor data quality would decrease the performance of such methods. Therefore, this section presents a comparison of two electricity consumption forecasts. The first one using data containing DQ problems and the second one relying on healthy data for training the algorithm.

#### 3.3.1 Random Forest as Regression Method for Forecasting

Several regression methods can be used to forecast electricity consumption. The models that result from the application of these methods can be used in numerous ways, as in demand-side management [10] or as a step in non-intrusive load monitoring evaluations. These models, when generated from healthy data can also be used to solve some data quality problems, such as in the reconstruction of profiles when there is a lack of data, or even in the identification of outliers and anomalies [11].

In this chapter, the random forest method [12] was applied as a regression/forecasting method. It is an ensemble machine-learning method for classification and regression, among other tasks. For classification problems, the output is the mode of all classes resulting from the individual trees. Meanwhile, for regression tasks, the result is the mean prediction of the outcomes from each tree in the forest [13]. In other words, this method creates several independent decision trees (a decision support tool that represents a set of choices in the graphical form of a tree) during the training phase, in a random way forming a forest. Every decision tree created is used in the result. Random decision forests correct for decision trees' habit of overfitting to their training set [14]. Figure 42 illustrates a schema of a decision tree, while Figure 43 shows the representation of the Random Forest algorithm.

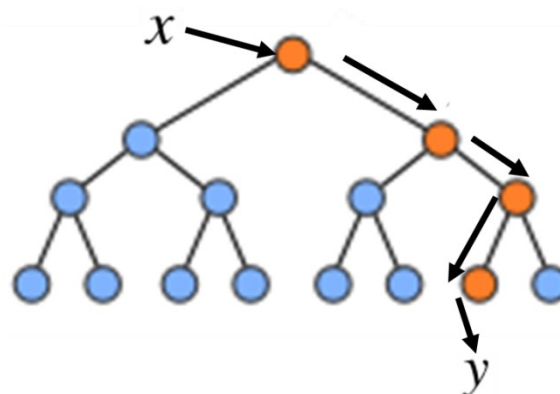


Figure 42 – Schema of the decision tree algorithm.

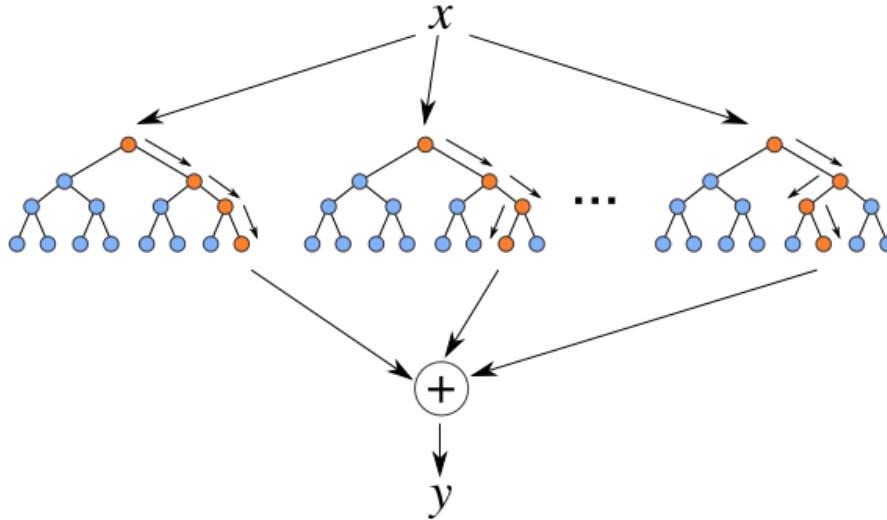


Figure 43 – Schema of the random forest algorithm. [15]

After the application of a forecast method, an assessment of its performance is needed. There are several metrics used to measure the global performance of a regression method. In this chapter, the mean absolute error (MAE) [16] and the mean absolute percentage error (MAPE) [17] were used. These metrics can be calculated using the following equations, respectively.

$$MAE = \frac{1}{N} \sum_{t=1}^N |Actual_t - Forecast_t| \quad \text{Equation 2}$$

$$MAPE = \frac{1}{N} \sum_{t=1}^N \frac{|Actual_t - Forecast_t|}{Actual_t} 100 \quad \text{Equation 3}$$

### 3.3.2 Example of the Impact of Poor Data Quality in Machine-learning Approaches

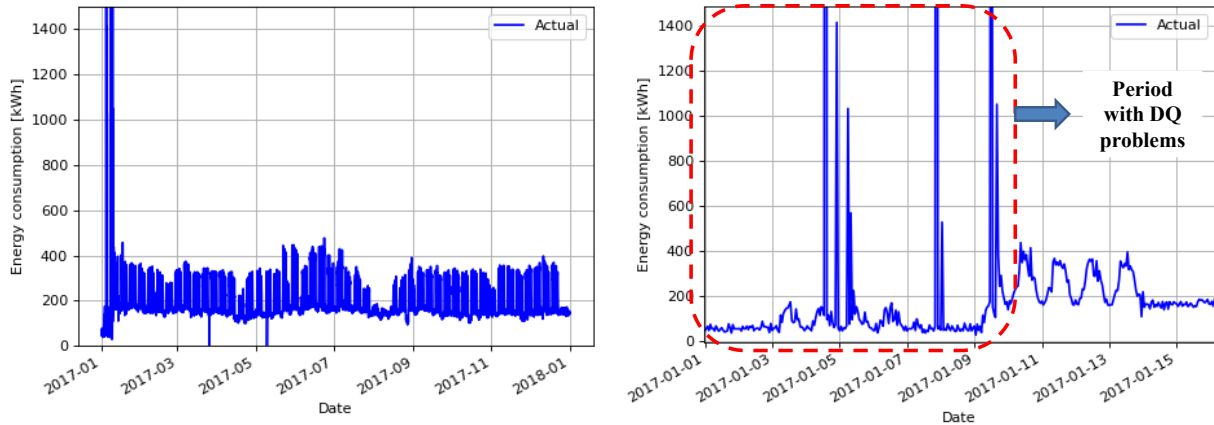
In order to show the potential impact of poor DQ in machine learning approaches, this section presents two examples of electricity consumption forecast. Firstly a forecast was performed using as training set data containing DQ problems. The second example, presents the same forecast using healthy data as training set.

A model for the GreEn-ER energy consumption was defined using random forest method, with 500 estimators. The data were resampled to 1 hour time interval, smoothening the series. For training of the algorithm, the following data features were used:

- External temperature;
- Average temperature of the day;

- Time of the day;
- Day of the year (with information of holidays and vacations).

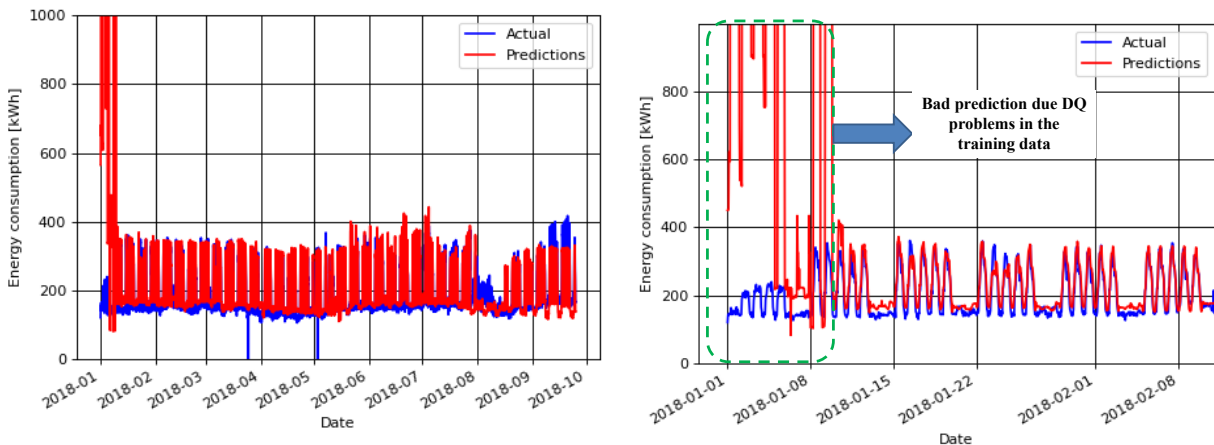
The training data set was defined as the global consumption of the building from 2017, using the constructed model to forecast the 2018 consumption until October. The results are presented in the following figures.



a) Load curve of GreEn-ER

b) Zoom in on the load curve highlighting a period with DQ problems

Figure 44 – Global load curve of GreEn-ER in 2017 used as training data.

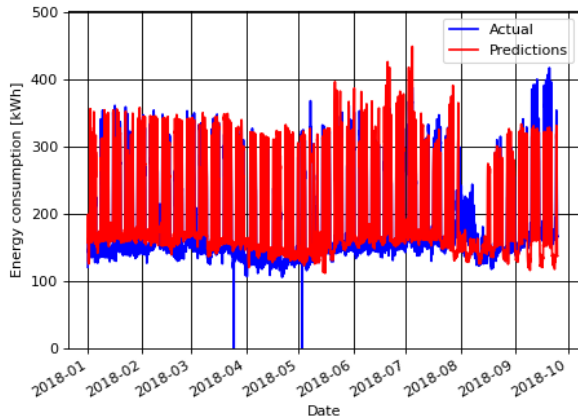


a) Results of the forecast

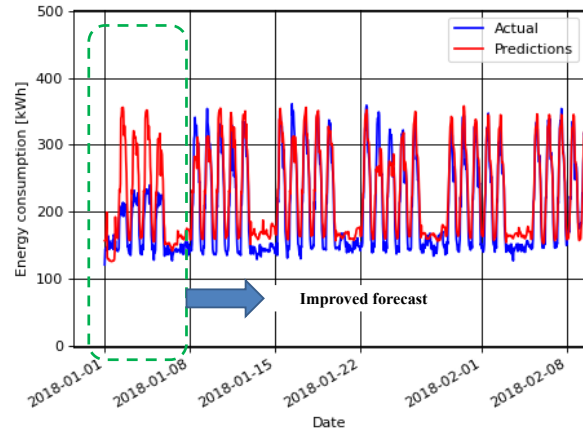
b) Zoom in on the results of the forecast

Figure 45 – Forecast results for 2018 using data with DQ problems as training data set.

The results presented in the previous figures show the poor prediction made by the random forest when using data with DQ problems as training set. On the other hand, when the training set does not contain, or contains less impacting, DQ problems, the forecast improves. This can be seen in Figure 46.



c) Results of the forecast



d) Zoom in on the results of the forecast

Figure 46 – Forecast results for 2018 using healthy data as training set.

Of course, the examples presented in this section are exaggerated. It is unlikely that anyone would perform a forecast with severe DQ problems in the training set, however, these examples illustrate how harmful poor data quality can be to machine-learning approaches.

### 3.4 Conclusions

This chapter have presented an overview on data quality issues and it influence on data-driven approaches, especially in the energy domain. By performing a consumption forecast for the GreEn-ER building, the present chapter exposed how harmful to a machine-learning technique could be a regardless selection of training dataset containing data quality issues.

Besides that, an assessment of the data quality of the GreEn-ER dataset was presented in terms of completeness and accuracy of a subdataset. Despite being relatively accurate in terms of energy consumption over the year of study (2017), the power retrieved from the energy consumption present inaccuracy. For instance, considering the exemplified subdataset, only 50% of the samples were under the 10% tolerance error when the sampling interval is 1 hour.

In addition, the dataset was also evaluated in terms of completeness. Missing data corresponded to 1.73%. However, this analysis is not complete, due the more than one way that this DQ issue is expressed in this dataset. To help to complete this task, an algorithm to identify the outliers following periods of communication issues was developed for the major GreEn-ER loads. This method is detailed in the next chapter.



## References

- [1] Pipino, L., Lee, Y., Wang, R. (2003). Data Quality Assessment. Communications of the ACM. 45, 2003, doi: 10.1145/505248.506010.
- [2] Caballero, I., Verbo, E., Calero, C., Piattini, M.. A data quality measurement information model based on ISO/IEC 15939, 2007, 393-408.
- [3] Redman, T.C., ed. Data Quality for the Information Age. Artech House: Boston, MA., 1996.
- [4] Ballou, D.P., Wang, R.Y., Pazer, H. and Tayi, G.K. Modeling information manufacturing systems to determine information product quality. Management Science 44, 4 (1998), 462–484.
- [5] International Organization for Standardization. ISO/IEC 25012:2008(E), Software Engineering—Software Product Quality Requirements and Evaluation (SQuaRE)—Data Quality Model; International Organization for Standardization: Geneva, Switzerland, 2008.
- [6] Wang, R.Y. and Strong, D.M. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems 12, 4 (1996), 5–34.
- [7] Ge, M., Helfert, M.: A framework to assess decision quality using information quality dimensions. In: Proceedings of the 11th International Conference on Information Quality, MIT, USA. pp. 455, 2016
- [8] Ge, M., Chren, S., Rossi, B., Pitner, T.. Data Quality Management Framework for Smart Grid Systems. 22nd International Conference on Business Information Systems (BIS2019), 2019, doi: 10.1007/978-3-030-20482-2\_24.
- [9] Chen, W., W, Zhou K, Yang S, Wu C. Data quality of electricity consumption data in a smart grid environment. Renew Sustain Energy Rev 2017;75:98. <https://doi.org/10.1016/j.rser.2016.10.054>.
- [10] Zhao, H.; Tang, Z. The review of demand side management and load forecasting in smart grid. In Proceedings of the 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China, 12–15 June 2016; pp. 625–629, doi:10.1109/WCICA.2016.7578513.
- [11] Vandeput, N. Data Science for Supply Chain Forecast; Independently published: Chicago, IL, USA. 2018, ISBN 978-1730969430.
- [12] Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282, doi:10.1109/ICDAR.1995.598994.
- [13] Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. Classification and Regression Trees; Wadsworth and Brooks: Monterey, CA, USA, 1984.
- [14] HARP. “Harp Random Forests”. Available online: <https://dsc-spidal.github.io/harp/docs/examples/rf/>. Accessed on 17 November 2021
- [15] Tierney, B. Random Forest Machine Learning in R, Python and SQL - Part 1. Toad World Blog. <https://blog.toadworld.com/2018/08/31/random-forest-machine-learning-in-r-python-and-sql-part-1> August 2018.
- [16] Sammut, C.; Webb, G.I. “Mean Absolute Error”. In Encyclopedia of Machine Learning; Springer: Boston, MA, USA, Available online: [https://doi.org/10.1007/978-0-387-30164-8\\_525](https://doi.org/10.1007/978-0-387-30164-8_525). Accessed on 17 November 2021.

- 
- [17] De Myttenaere, A.; Golden, B.; Le Grand, B.; Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing* 2016, 192, 38–48, <https://doi.org/10.1016/j.neucom.2015.12.114>.



## 4 Outlier Detection in Buildings' Power Consumption Data Using Forecast Error

Chapter 3 discussed the issues that poor data quality, especially the presence of outliers, can create into data-driven approaches such as energy efficiency and energy sobriety analysis, or even machine-learning applications. To work around these issues, Chapter 4 presents an algorithm that combines machine learning techniques and classic statistical methods to detect outliers present within the GreEn-ER dataset. Considering both cases tested, this combination has presented the best F-score, but it was not perfect. Hence, a human-in-the-loop approach is still needed, with the forecast error outlier detection method pointing out potential outliers and a human agent validating them. Thus, the effort would be less costly with the application of the method presented in this chapter. Later, this technique was employed to assess the completeness of the TGBT1 load.

### Chapter Contents

4.1 Forecast Error Method Overview .....	84
4.2 Statistical Methods for Outlier Detection .....	85
4.3 Results and Analysis.....	90
4.4 Assessing Completeness of GreEn-ER loads.....	99
4.5 Conclusions.....	101

## 4.1 Forecast Error Method Overview

As stated in previous chapters, the use of machine-learning techniques is increasing in the energy sector. Moreover, the previous chapter highlighted problems that poor data quality may cause into data-driven applications, such as energy efficiency and energy sobriety analysis, or even machine-learning applications. For example, the anomaly detection using machine-learning techniques can help to identify unusual energy consumption of assets [2–4] and detect equipment faults [5]. The importance of detecting outliers, either if they indicate unusual energy consumption or problems in the metering system encouraged the focus on this data quality dimension on the GreEn-ER data.

Several methods for the detection of outliers have been used in recent times. Classic statistical methods, such as the three-sigma rule [6] and the boxplot method [7], have been highly used. However, these techniques assume a symmetrical data distribution and the performance of these techniques is highly dependent on this feature, which is commonly unknown for power consumption data.

To work around the issue of unknown data distribution, researchers have used regression-based methods to tackle this problem. The first step, called the training phase, comprises the definition of a regression model that fits the data. After the construction of the model, every data sample is compared with the model instances in the test phase [8]. A data point is labeled as an outlier if a remarkable deviation occurs between the actual value and its expected value produced by the regression model [9]. Several techniques were used to detect outliers using regression methods. For example, in [10] the author used linear regression to detect outliers and in [11] an auto regressive moving average (ARMA) was used as the regression technique.

Therefore, this chapter aims at the development and application of a hybrid method, combining regression techniques and classical statistic outlier methods focusing on detecting outliers of a dataset that contains measurements of electrical energy consumption of a tertiary building. The random forest [12] method as a regression technique to construct a model was used in this chapter. Afterwards, all measured samples were compared with the model instances, resulting in an error. The statistical outlier detection methods were then implemented to search high error values in order to classify them as potential outliers. This combination is called the forecast error method.

The construction of a predictive model of energy consumption in a building can be of great importance for energy managers. Through these models, it is possible to plan from the short term optimization of energy consumption costs to the allocation of assets in case of

preventive maintenance with low impact on the building's normal activity. In addition, the implementation of an algorithm that can detect anomalies integrated into a building management system can facilitate the identification of potential energy consumption reduction or even the need to perform corrective maintenance on an asset that may present a defect in real time.

The following section exposes the statistical outlier detection methods employed in this chapter. Afterwards, the regression (forecast) method employed and the error metrics that can be used to assess the performance of this method are briefly introduced. The combination of these techniques is applied in two datasets. In the first one, called "adapted data", twelve outliers were manually introduced in healthy synthetic electricity consumption data, adapted from the GreEn-ER data. The second one consists in "real data" measurements of the electricity consumption, with outliers generated by problems inherent to buildings' metering systems. The results show that the combination of a regression technique and the adjusted boxplot method [13] presents the better performance compared with the other methods when searching outliers in the tested datasets.

## 4.2 Statistical Methods for Outlier Detection

An outlier is an observation that deviates so much from other observations that it rises suspicions that it was generated by a different mechanism [14]. This type of sample can indicate malfunctioning of the metering system or even in a load itself. In addition, if this data quality problem is too persistent it can affect the accuracy of eventual machine-learning algorithms using this dataset. Therefore, its identification and correction are important steps in the data pre-processing.

Standard outlier detection consists of two main components. The first one is calculating an outlier score for every data instance. The outlier score can be the value itself or even the difference between the value and its prediction [15], considering that the prediction model was generated based on healthy data. Other formulations, such as the local outlier factor [16], calculate this score by comparing the value to its k-nearest neighbors in a feature space. The second component is thresholding the outlier scores by the application of some statistical methods. This step decides how highest scoring points are labelled as outliers.

In this chapter, several statistical methods for thresholding were applied: the three sigma rule [6], the median absolute deviation [17], the original boxplot [7], the skewed boxplot [18] and finally the adjusted boxplot [13]. Each of these methods is detailed in the

following sections.

### 4.2.1 Three-Sigma Rule

The three-sigma rule is a simple and heuristic method for outlier detection [6]. In a normal distribution, the probability of a sample to be within the range between  $\mu \pm 3\sigma$ , where  $\mu$  is the mean and  $\sigma$  is the standard deviation (STD), is 99.7%, as shown in Figure 47.

Therefore, the upper and lower bounds that defines if a value is an outlier or not, can be calculated by applying the following equations in which UPB represents the upper bound and LWB the lower bound. Samples that are higher than the upper bound, or lower than the lower bound, are potential outliers.

$$UPB = \mu + 3 * \sigma \quad \text{Equation 4}$$

$$LWB = \mu - 3 * \sigma \quad \text{Equation 5}$$

Since the three-sigma rule is based on the mean value and the standard deviation, this method is sensitive to the presence of extreme outliers.

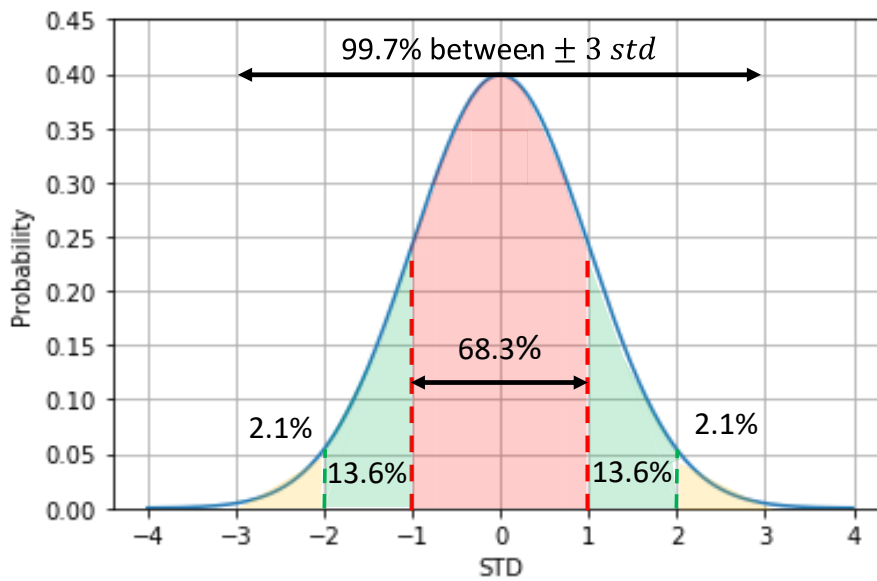


Figure 47 – Percentages in normal distribution between standard deviations. Based on Straker [19].

#### 4.2.2 Mean Absolute Deviation (MAD)

Because of its sensitivity to the presence of outliers, the mean value is not the most suitable measure of central tendency to be used in the outlier detection. The median value, another measure of central tendency, is more adapted to this task due to its insensitivity to the existence of outliers in the dataset. The median is defined as the value associated with the mean rank after sorting the data ascendingly.

The median absolute deviation [20] is then defined as the median of the absolute deviation from the median, and can be described as follows:

$$MAD = b * med(|x_i - med(X)|) \quad \text{Equation 6}$$

In this equation,  $b$  is a constant, suggested as 1.4826,  $x_i$  represents each sample and  $X$  is the vector that contains all samples. The upper (*UPBM*) and lower (*LWBM*) bounds can be calculated by the application of the following equations:

$$UPBM = med(X) + 3 * MAD \quad \text{Equation 7}$$

$$LWBM = med(X) - 3 * MAD \quad \text{Equation 8}$$

#### 4.2.3 Boxplot

The modern boxplot, described in more detail by Tukey [7], is a graphical method for detecting potential outliers through a box and whiskers plot with restrictions on the data used [21]. In order to provide a robust measurement of the data series, the boxplot uses some characteristic values of the series, such as the median and the values of the first (25%) and the third (75%) quartiles. Using these quartile values, the interquartile interval is calculated applying Equation 9, in which *IQR* represents the interquartile range and  $Q_3$  and  $Q_1$  represent the values of the first and third quartiles, respectively.

$$IQR = Q_3 - Q_1 \quad \text{Equation 9}$$

Based on the values of the quartiles ( $Q_3$  and  $Q_1$ ) and the interquartile range (*IQR*) it is then possible to determine the upper (*UPBB*) and lower (*LWBB*) bounds for the boxplot method by applying Equation 10 and Equation 11. Values located beyond these limits are considered potential outliers. In his work, Tukey [7] proposed that  $K = 1.5$  indicates potential mild outliers and  $K = 3$  classifies the sample as a potential extreme outlier.

$$UPBB = Q_3 + K * IQR \quad \text{Equation 10}$$

$$LWBB = Q_1 - K * IQR \quad \text{Equation 11}$$



When the data distribution follows a normal characteristic, this method includes 99.3% of the data within its limits [21] when  $K = 1.5$ , as can be observed in Figure 48.

Since the Boxplot method uses the positional values of the samples in the series, and not their values directly, this method is less sensitive to the presence of extreme outliers.

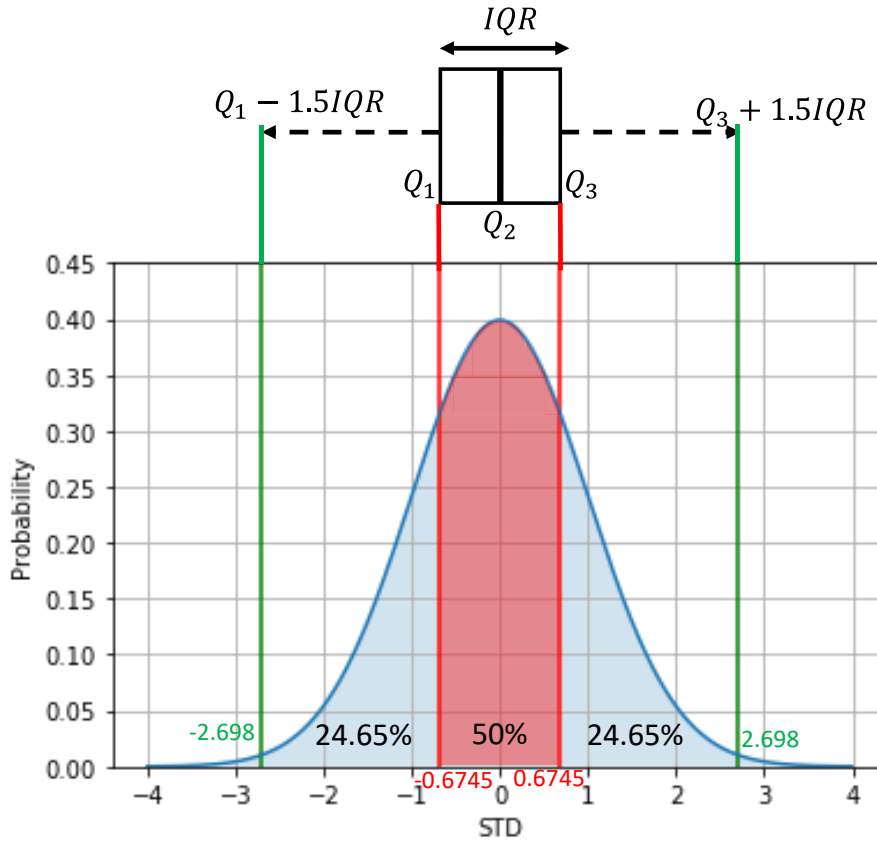


Figure 48 – Example of a box-and-whisker plot for a normal distribution. Based on Olano et al. [22].

#### 4.2.4 Skewed Boxplot

The boxplot method, described in the previous section, is better suited for the detection of outliers in a dataset whose distribution is symmetric. When the distribution is skewed, some samples that exceed the upper and lower bounds defined by that method may be misclassified as outliers [23]. For this reason, a correction is necessary in the calculation method of the upper and lower bounds.

There are some ways to adjust the boundaries considering asymmetrical data. In 1990, Kimber [18] proposed a method to consider the skewness of distribution in the search for outliers. The following equations define the upper ( $UPBS$ ) and lower ( $LWBS$ ) bounds used in this method. In these equations,  $SIQR_U$  is the upper interquartile range,  $SIQR_L$  is the lower interquartile range and  $Q_2$  represents the median of the evaluated series (or the second quartile).

$$SIQR_U = Q_3 - Q_2 \quad \text{Equation 12}$$

$$UPBS = Q_3 + 3 * SIQR_U \quad \text{Equation 13}$$

$$SIQR_L = Q_2 - Q_1 \quad \text{Equation 14}$$

$$LWBS = Q_1 - 3 * SIQR_L \quad \text{Equation 15}$$

#### 4.2.5 Adjusted Boxplot

Another method to consider the skewness of a data distribution and to adjust the boundaries is the adjusted boxplot, proposed by Hubert and Vandervieren [13]. In this method, the medcouple ( $MC$ ), proposed by Brys et al. [23], is used as a magnitude to measure the asymmetry of the evaluated series. This quantity is defined by the following equations.

$$MC = med(h(x_i, x_j)) \quad \text{Equation 16}$$

$$h(x_i, x_j) = \frac{(x_j - med(X)) - (med(X) - x_i)}{x_j - x_i} \quad \text{Equation 17}$$

$$x_i \leq med(X) \leq x_j \quad \text{Equation 18}$$

In these equations,  $x_i$  represents the samples of the series smaller than, or equal to, the median and  $x_j$  the samples larger than, or equal to, the median. Thus, to adjust the boxplot method according to the asymmetry of the evaluated series, the medcouple is incorporated in the calculation of the upper and lower bounds. For left-skewed data, with negative medcouple, the limits are calculated as shown in the following equations.

$$UPBA = Q_3 + 1.5 * e^{4MC} * IQR \quad \text{Equation 19}$$

$$LWBA = Q_1 - 1.5 * e^{-3MC} * IQR \quad \text{Equation 20}$$

In which  $UPBA$  and  $LWBA$  represent the upper and the lower bounds, respectively. For right-skewed data with positive medcouple, the following equations are used.

$$UPBA = Q_3 + 1.5 * e^{3MC} * IQR \quad \text{Equation 21}$$

$$LWBA = Q_1 - 1.5 * e^{-4MC} * IQR \quad \text{Equation 22}$$

#### 4.2.6 Error Metrics for Classification

Labeling samples as outliers or normal samples is a classification problem. Various are the metrics to assess the performance of the algorithms used to tackle this kind of problem. In the present chapter, the concepts of precision, recall [24] and the F-score (or F1) [25] are applied. These metrics are defined by the following equations.

$$Precision = \frac{TP}{TP + FP} \quad \text{Equation 23}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{Equation 24}$$

$$F1 = 2 * \left( \frac{Precision * Recall}{Precision + Recall} \right) \quad \text{Equation 25}$$

In which  $TP$  is the number of true positives classifications (actual outliers detected),  $FP$  the number of false positives classifications (normal samples misclassified as outliers), and  $FN$  is the number of false negatives (undetected outliers).

In the context of the outlier identification task, precision indicates the proportion of actual outliers identified among all potential outliers flagged by the search method. On the other hand, recall is related to the number of outliers not flagged by the algorithm. The F-score uses the harmonic mean between both to evaluate the global accuracy of the method.

### 4.3 Results and Analysis

As mentioned in the previous section, the outlier scores can be calculated by several approaches. In this chapter, the value itself and the difference between the actual value and its prediction were tested. To forecast, the random forest, presented in the previous chapter was used as regression method.

This section presents the results obtained by applying the forecast error method in the search for outliers in power consumption data of a tertiary building. Firstly, the regression methods results are shown, quantifying their performance through error metrics. Afterward, using these regressions, the forecast error method was applied. For comparison, the classic statistic methods for outlier detection were also applied so that the results from both techniques are presented. The code used to perform these tasks was developed in python language in a Jupyter Notebook, available in an online open repository [31].

The data used in this chapter were adapted from the dataset available for downloading at the open science platform Mendeley Data [32], and presented in Chapter 2. The data were then resampled as the hourly consumption, resulting in 8760 samples.

Two different data series were used to test the forecast error method. Firstly, some known outliers were inserted in synthetic healthy data, without outliers or any other data quality problem in order to establish a benchmark. This series was called adapted data. In a second stage, the technique was employed in a data series without any pre-treatment regarding data quality. Because of the presence of outliers in the beginning of the series, the

data was reversed, in order to use “past” data as training. This second dataset was called real data.

### 4.3.1 Adapted Data

The synthetic data, free of data quality problems, are illustrated in Figure 49. This dataset was created to simulate the behavior of the GreEn-ER building, and it was based on its own electricity consumption.

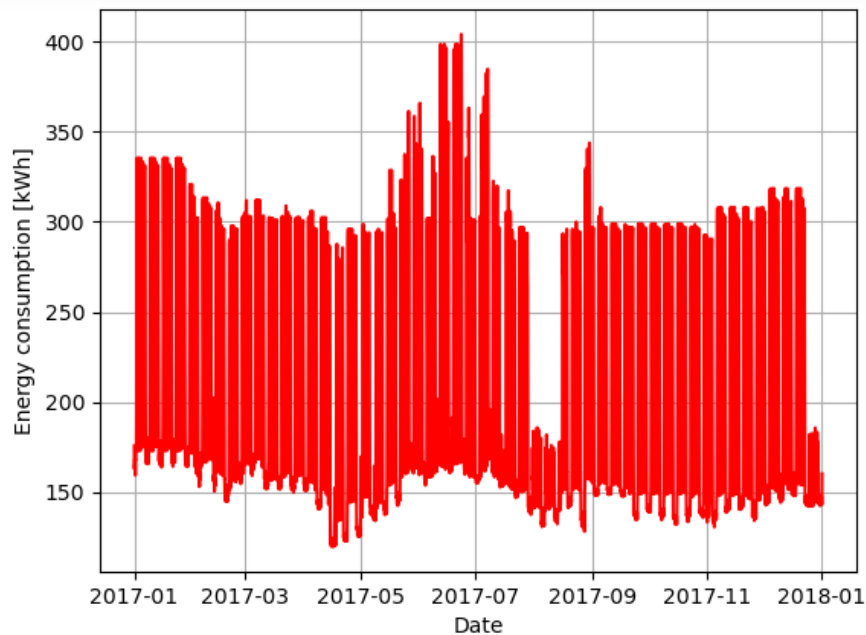


Figure 49 – Synthetic GreEn-ER global power consumption.

From the data exposed in the previous figure, it is possible to notice different consumption patterns for different periods. It can be seen that the periods of higher consumption match with those of higher occupation, during daytime on the weekdays. Outside these periods, during nighttime on the weekdays, weekends, holidays and vacations, the consumption reduces drastically. In addition, it is possible to notice a relation with the temperature since the highest consumption occurs during summer.

In order to test the outlier detection techniques, twelve outliers, both upper and lower, were manually introduced in the series presented in Figure 49, resulting in the dataset presented graphically in Figure 50. The information of these samples is shown in Table 16, and some of these outliers are highlighted in Figure 50 too. This information is then used as ground truth and compared with the results obtained to assess the classification of each sample in true positive, false positive and false negative.

Table 16 – Outliers inserted in the data series.

Outlier Index	Timestamp	Day of the Week	Holiday or Vacation	Value [kWh]	Type of Outlier
1	22/10/2017 09:00	Sunday	No	390	Upper
2	24/10/2017 10:00	Tuesday	No	430	Upper
3	26/10/2017 22:00	Thursday	No	87	Lower
4	29/10/2017 10:00	Sunday	No	95	Lower
5	04/11/2017 22:00	Saturday	No	405	Upper
6	07/11/2017 23:00	Tuesday	No	400	Upper
7	10/11/2017 13:00	Friday	No	93	Lower
8	12/11/2017 03:00	Sunday	No	120	Lower
9	26/12/2017 16:00	Tuesday	Yes	110	Lower
10	28/12/2017 14:00	Thursday	Yes	350	Upper
11	29/12/2017 05:00	Friday	Yes	105	Lower
12	30/12/2017 21:00	Saturday	Yes	375	Upper

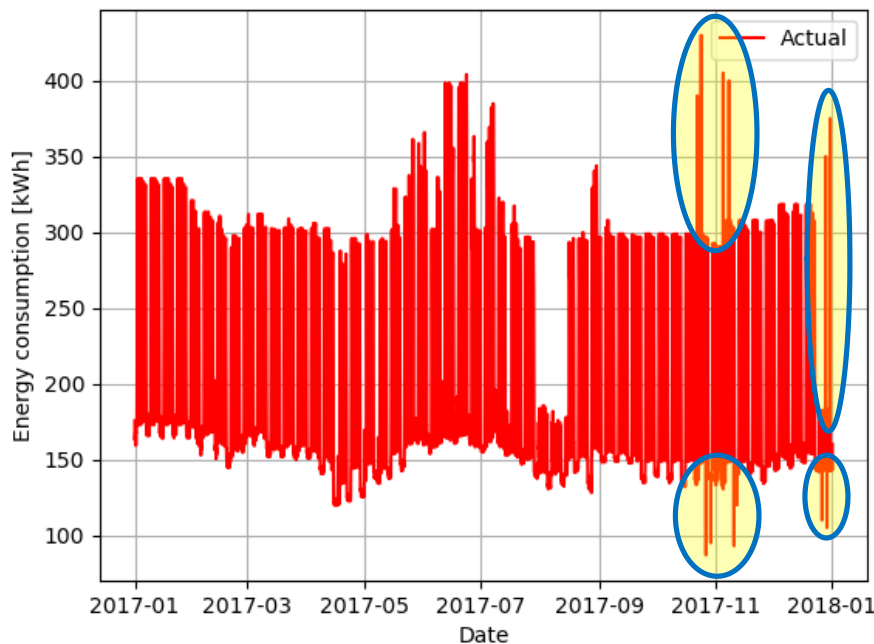


Figure 50 – Synthetic GreEn-ER global power consumption with inserted outliers.

#### 4.3.1.1 Regression Methods Results

In order to find a model for the GreEn-ER energy consumption, the random forest method was applied using the data exposed in the previous section as the regression technique. For training the algorithm, the same features already used in the previous chapter (external temperature, average temperature of the day, time of the day and day of the year (with information of holidays and vacations)) were employed.

The training dataset was defined as 80% of the data, from the beginning of the year until mid-October. All the outliers inserted in this dataset are concentrated beyond this period. These data quality problems make it difficult to assess the performance of the regressor only in test time interval because of their effect in the statistical variables (mean, median, standard deviation) used also to detect these abnormal samples. Because of that, the regressor performance was evaluated in two conditions. The first one considers the whole year, including the weeks with data quality problems and the training phase. The second one considers the period of the year complementary to the training phase. Figure 51 details, as an example, the results obtained with the application of the random forest method, using five hundred estimators as parameter. At the same time, Table 17 quantifies the performance of these regressions with the two conditions cited above.

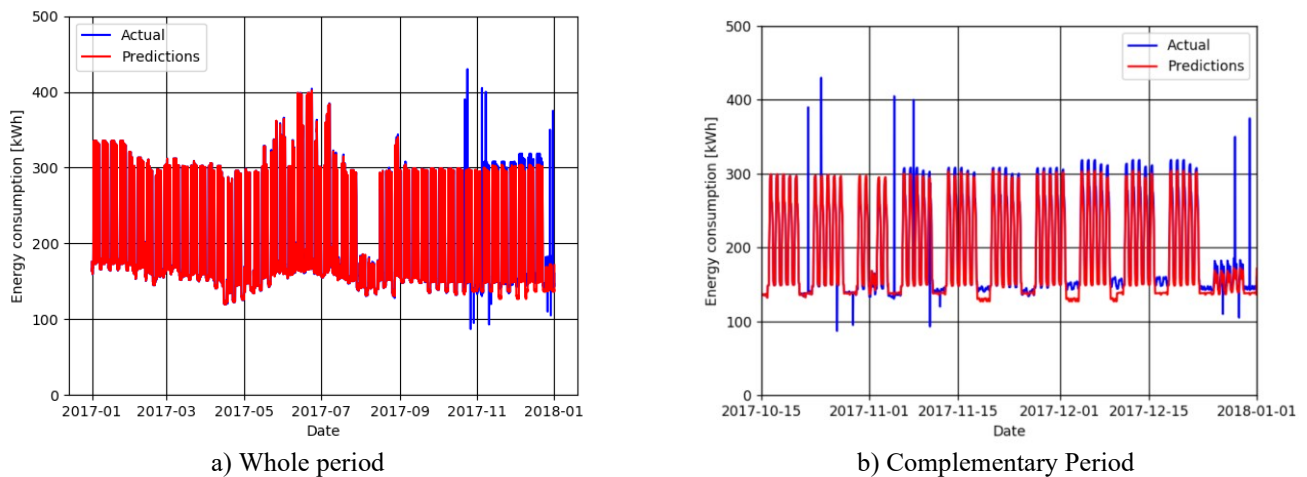


Figure 51 – Regression results using the random forest algorithm on adapted data.

Table 17 – Performance of the regression methods on the adapted data.

Error Metric	Period	
	Complete	Complementary Period
MAE	1.98	8.36
MAPE	1%	4.27%

Observing Figure 51b, the predictor presents satisfactory results. Although it underestimated the power on the weekends, the predictor was able to detect the daily and weekly patterns and even during the holidays, resulting, on average, in less than a 5% error.

Regarding the regression, the most important features are the hour of the day, reproducing the daily pattern of the consumption, and the day of the week, reproducing the weekly shape of the load curve. The holiday feature also plays an important role in the performance of the regressor. The other features would be more important if more than a year’s worth of data were available. For instance, the external temperature would improve the

regression inserting the season component, such as the difference between the days from summer and winter. However, with one year’s worth of data, and the choice of taking 80% of the series as training, this component is not important. These features were maintained in the model with the objective to improve the model in a future real time application, when more than a year would be available. Figure 52 shows the feature importance of the regression made for the adapted data.

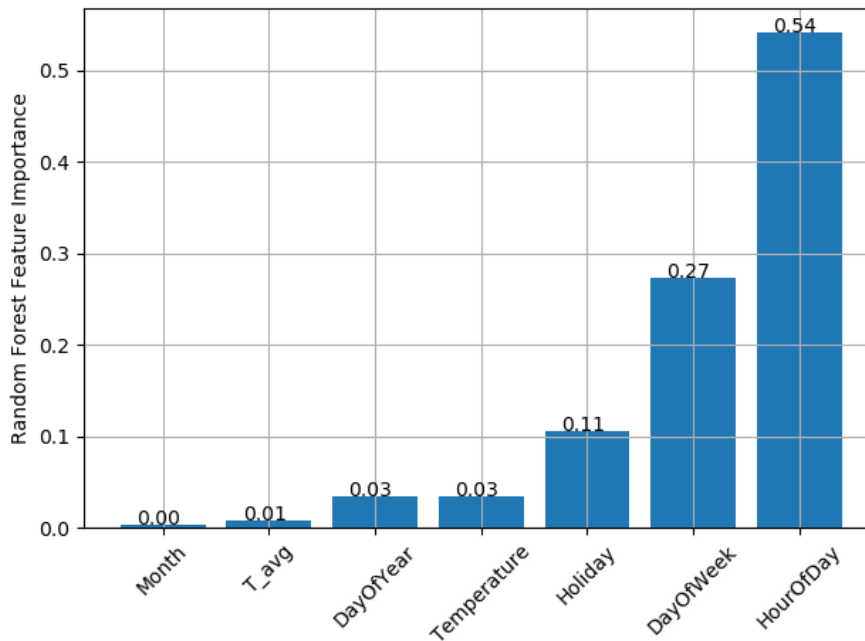


Figure 52 – Data features’ importance in the random forest regression for the adapted data series.

#### 4.3.1.2 Outlier Detection

In order to detect the outliers inserted in the data series, two strategies were applied. Primarily, a global search employing the statistical methods on the power consumption data was performed. Afterward, they were used to search outliers via the forecast error. Twelve outliers were manually inserted, six of them were upper outliers and the other six, lower, as shown in the previous section.

In the global search strategy, the search for outliers was performed only once. In this way, all information available is used, and the outliers are assumed to have any, or low, influence on the average value, the standard deviation, or even on the quartile values. Therefore, the global search was performed using the three-sigma rule, the boxplot, the skewed boxplot and the adjusted boxplot methods. The results are shown in Table 18. In this table, the column Potential Outliers Detected indicates the number of samples flagged out as outliers by each method. In the True Positives column, there are the number of actual outliers

detected, while in the False Negatives column, the number of undetected outliers is presented. Furthermore, in the False Positives column, the number of normal samples misclassified as outliers are shown. Therefore, the sum of the true positives and the false negatives should be equal to the number of outliers present in the dataset, in this case, twelve. The sum of the true positives and false positives is equal to the potential outliers detected and the sum of both false positives and negatives gives the total of samples misclassified by each method.

Table 18 – Number of outliers found in the global search by each method on adapted data.

Method	Potential Outliers Detected	True Positives	False Negatives	False Positives	Total Misclassifications	Precision	Recall	F-Score
3 Sigma	11	3	9	8	17	0,273	0,25	0,261
MAD	808	6	6	802	808	0,007	0,5	0,015
Boxplot	0	0	12	0	12	0	0	0
Skewed Boxplot	1	1	11	0	11	1	0,083	0,154
Adjusted Boxplot	82	6	6	76	82	0,073	0,5	0,128

The results indicate that the MAD and the adjusted boxplot were the most successful methods in detecting outliers, having found half of them; however, they still misclassified several other samples, reducing their precision. Thus, even detecting some outliers, their poor recall, with several false positive samples classified as outliers, show that these methods alone are not the best suitable to detect outliers, especially local ones, such as those inserted in this dataset.

As the classical statistical methods failed to detect several outliers in the study dataset, the forecast error method, which compares the results of previous regression models with measurements was employed. The statistical methods for outlier detection are then applied on the resulting error. Table 19 shows the number of outliers detected by each method considering the deviation between the actual values and the predictions.

Table 19 – Number of outliers found by the Forecast Error method applied to the random forest forecasts on the adapted data.

Method	Potential Outliers Detected	True Positives	False Negatives	False Positives	Total Misclassifications	Precision	Recall	F-Score
3 Sigma	23	11	1	12	13	0.478	0.917	0.628
MAD	2136	12	0	2124	2124	0.005	1	0.011
Boxplot	1214	12	0	1202	1202	0.01	1	0.02
Skewed boxplot	1301	12	0	1289	1289	0.009	1	0.018
Adjusted boxplot	20	11	1	9	10	0.55	0.917	0.688

The results presented in Table 19 indicate that all the methods were able to detect most of the outliers inserted in the dataset, using the forecast error. However, the poor precision of



the MAD, the boxplot, and the skewed boxplot misclassifying several samples indicates that they are not well suitable for this task in this dataset. The other two, three-sigma rule and adjusted boxplot, perform better and similarly, with a small advantage for the adjusted boxplot.

### 4.3.2 Real Data

The forecast error method was also tested in a dataset, available for downloading at the open science platform Mendeley Data [32], with no pre-treatment regarding data quality. This dataset was extracted directly from the GreEn-ER Building Management System and contains several problems of data quality, inherent to this type of monitoring. Figure 53 illustrates the power consumption data of the GreEn-ER building in which it is possible to visualize, for example, some outliers, values that extrapolate the scale of the graph, at the end of the year. In that period, both upper and lower outliers can be seen. A human agent looked through all samples and classified them into normal samples and upper (values higher than the normal instances) and lower (values lower than the normal instances) outliers, establishing the ground truth to which the results are compared to determine the true positives, false positives, and false negatives. Table 20 shows the type and the number of outliers found by the human agent.

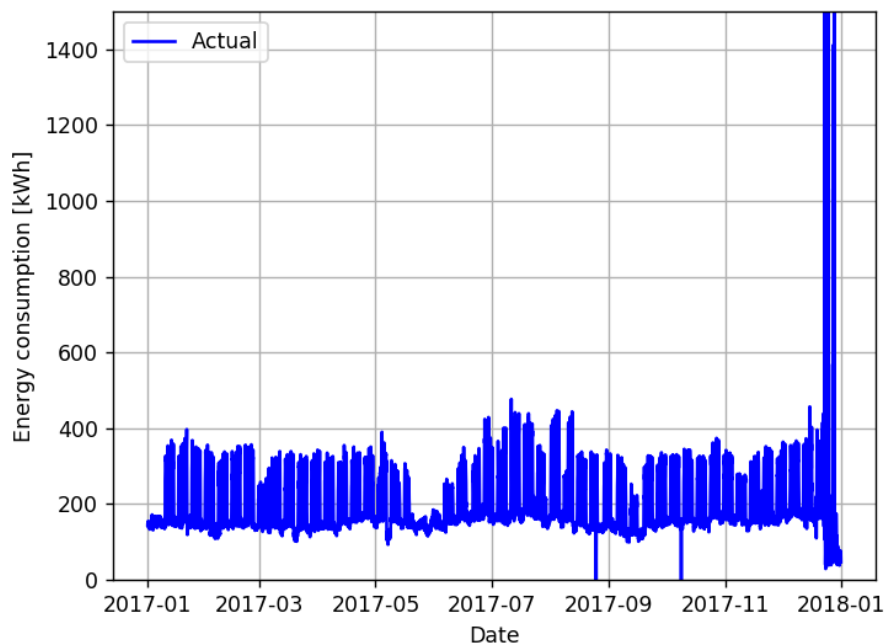


Figure 53 – Real GreEn-ER global power consumption with inserted outliers.

Table 20 – Number of outliers on real data found manually.

Upper Outliers	Lower Outliers	Total Outliers
8	204	212

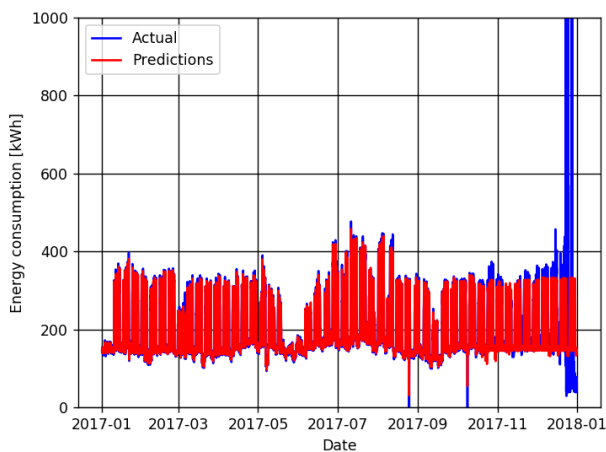
### 4.3.2.1 Regression Methods Results

The procedure shown in the previous section was applied to the real data series. The random forest method was employed as the regression technique, with unmodified parameters and the results obtained are presented in Figure 54. The performance of the regression is quantified in Table 21.

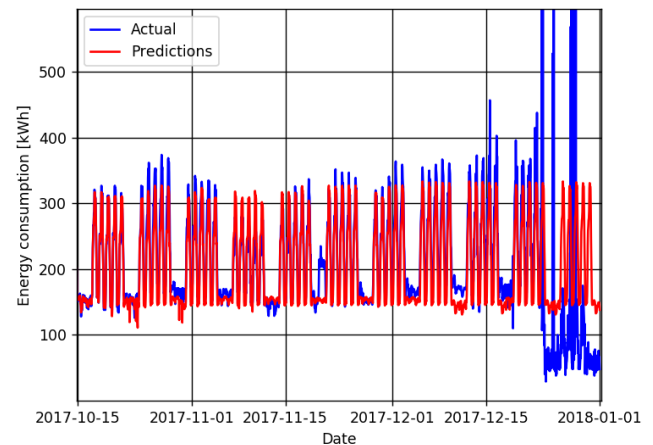
Table 21 – Performance of the regression methods on the Real Data.

Error Metric	Period	
	Complete	Complementary Period <sup>1</sup>
MAE	12.08	19.63
MAPE	6.82%	8.95%

<sup>1</sup> Excluding last week.



a) Regression results and actual data during the whole year



b) Regression results and actual data during the period complementary to the training phase.

Figure 54 – Regression results using the random forest algorithm on real data

Considering the complementary period, in Figure 54b, it can be seen that the predictor was able to reconstruct the daily and weekly patterns of the building consumption. The results are satisfactory, with less than 8% error on average, excluding the last week which contains numerous severe data quality problems. These anomalies are the ones that need to be pointed out, so the imperfection of the predictor is expected.

Regarding the importance of the features of the regression, similar results to the adapted data were obtained. Figure 55 shows the importance of each feature in the regression of the real data series.

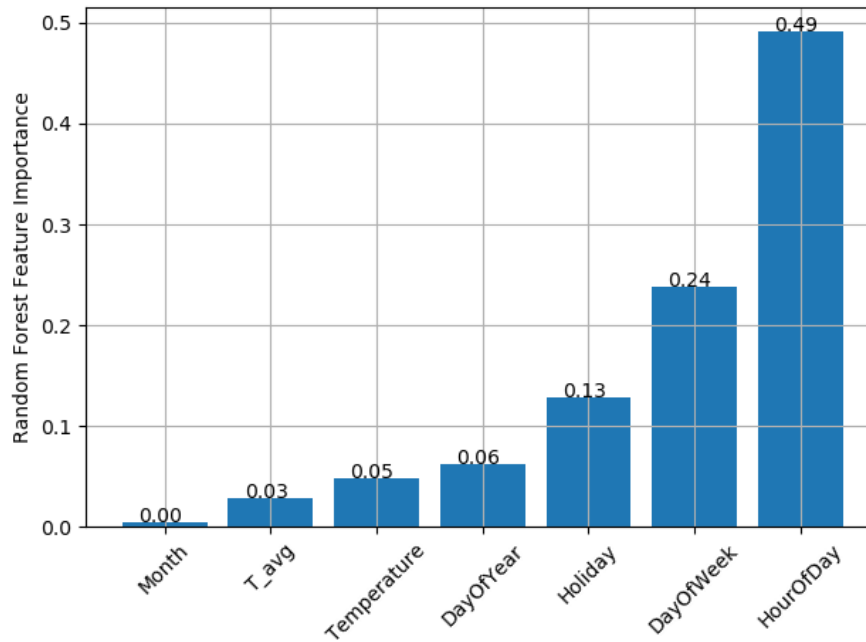


Figure 55 – Data features importance in the random forest regression of the real data series.

#### 4.3.2.2 Outlier Detection

As previously shown in Table 20, the outliers present in the series were manually classified to establish a benchmark for comparing the performance of the outlier detection algorithms. Two hundred and twelve outliers were found, eight of which are upper outliers and the other two hundred and four, lower.

The global search was performed using the three-sigma rule, the boxplot, the skewed boxplot and the adjusted boxplot methods. The results are shown in Table 22.

Table 22 – Outliers found in the global search by each method on real data.

Method	Potential Outliers Detected	True Positives	False Negatives	False Positives	Total Misclassifications	Precision	Recall	F-Score
3 Sigma	6	6	206	0	206	1	0.028	0.055
MAD	931	11	201	920	1121	0.012	0.052	0.019
Boxplot	6	6	206	0	206	1	0.028	0.055
Skewed boxplot	152	151	61	1	62	0.993	0.712	0.830
Adjusted boxplot	271	172	40	99	139	0.635	0.811	0.712

The presented results indicate that none of the tested methods were able to detect all the outliers. Furthermore, although the adjusted boxplot has initially pointed out more outliers than actually exist, it failed to detect forty outliers, and misclassified ninety-nine normal samples as abnormal data instances. Therefore, the results corroborate that those classical statistical methods applied to the value itself are not suitable to detect outliers, especially for local ones, such as the lower outliers present in this dataset. As shown in the previous section,

the forecast error method was also employed in the real data.

Although they found all outliers of the dataset, the results presented in Table 23 corroborate the fact that the MAD, the boxplot, and the skewed boxplot are not the most adapted methods to detect outliers using the forecast error in this dataset, as they misclassified several samples as outliers. Furthermore, the three-sigma rule failed to detect most of the outliers, flagging only the obvious upper outliers. Finally, the adjusted boxplot performed better, but still misclassified some samples. This method was able to detect 192 out of 212 outliers and misclassified another 13 samples as outliers, resulting in 33 misclassifications. This better performance of the adjusted boxplot can be seen by observing the F-score. While the MAD, the boxplot, and the skewed boxplot all have 1 recall, meaning that they found all the outliers (zero false negatives), their misclassification is costly as shown in their poor precision. This affects the F-score, decreasing its value. On the other hand, the adjusted boxplot has presented the best compromise between the precision and the recall, resulting in both metrics to be higher than 0.90.

Table 23 – Number of outliers found by the forecast error method applied to the random forest forecasts on the real data.

Method	Potential Outliers Detected	True Positives	False Negatives	False Positives	Total Misclassifications	Precision	Recall	F-Score
3 Sigma	6	6	206	0	206	1	0,028	0,055
MAD	1458	212	0	1246	1246	0,145	1	0,254
Boxplot	860	212	0	648	648	0,247	1	0,396
Skewed boxplot	1056	212	0	844	844	0,201	1	0,334
Adjusted boxplot	205	192	20	13	33	0,937	0,906	0,921

#### 4.4 Assessing Completeness of GreEn-ER loads

Chapter 3 has presented analyses of data quality regarding the GreEn-ER dataset, especially in terms of accuracy and completeness. However, the completeness analysis was incomplete. In normal operation periods, energy meters send the consumed energy value frequently, depending on the sampling interval. Nevertheless, it is not unusual for failures in communication between the meters and the data storage system to occur. In such cases, energy meters store the cumulative energy consumption, while the BMS keeps recording the last value received. Once the communication is restored, the meter sends the value corresponding to the energy consumption relative to the failure period. That is the cause of discontinuities in the energy consumption graph, shown in the Figure 37. In terms of power, this data quality issue is usually expressed by a sequence of zero power followed by a peak

value, which can also be seen in Figure 37.

However, it is not trivial to distinguish true zero consumption from the zeros caused by this issue, especially in the case of small loads or during a small duration of the issue, in which the outliers can be confused as normal values.

Hence, the algorithm presented in this chapter was employed to assess the completeness of the GreEn-ER dataset. To this, the TGBT1 load was selected as subdataset. Figure 56 presents the TGBT1's load curve.

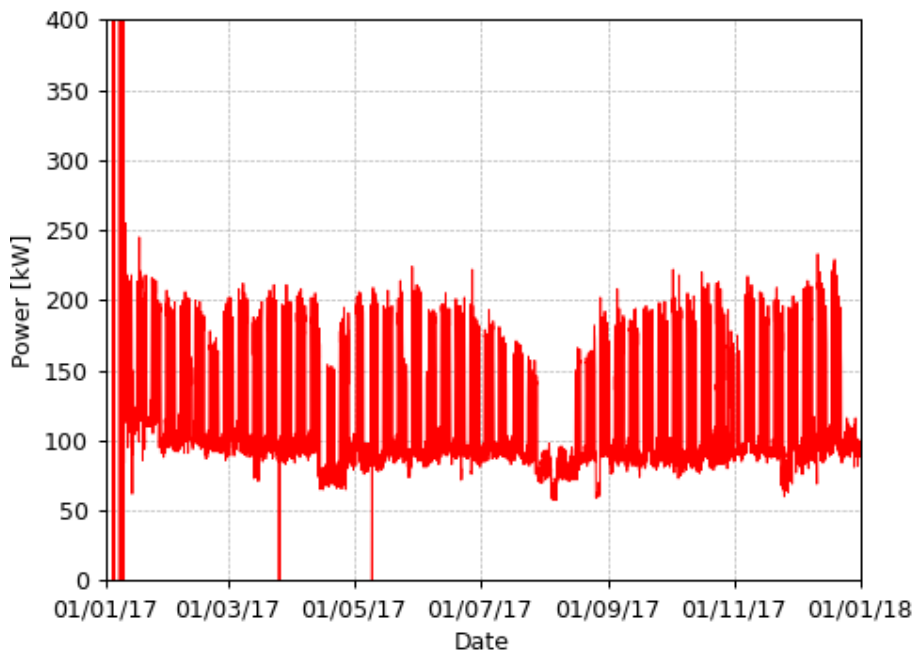


Figure 56 – TGBT1 load curve.

The algorithm has detected 219 potential outliers, comprehending both upper and lower outliers. The outlier values are then sorted into higher and lower than the power average value. To assess the completeness in this case, it is necessary to count how big the sequence of zeros immediately before the potential upper outlier is. In the case of the TGBT1, 201 samples equal to zero were found immediately before the upper outliers, indicating that that is the number of samples missing due communication issues between the meter and the data storage system. This analysis was performed considering 1 hour as sampling interval, which means that the completeness due these issues is 97.71%. The lack of completeness in this case represents more than 1 week worth of information lost due to the issues mentioned in this section.

## 4.5 Conclusions

This chapter aimed to employ a hybrid method, called forecast error, to detect outliers in the power consumption of a tertiary building. This method combines regression methods with statistical outlier detection techniques. The random forest algorithm was used as the regression method and the three-sigma rule, the median absolute deviation, the boxplot, the skewed boxplot, and the adjusted boxplot were chosen as outlier detection techniques. In a global search, using only the statistical methods to the data instances themselves, none of them has presented the expected performance. On the other hand, when the adjusted boxplot was applied to the forecast error (difference between the actual measurement and the forecast) better performance was obtained. Considering both datasets tested, this combination has presented the best F-score (higher than 0.90 in the real data dataset), but it was not perfect. Hence, a human-in-the-loop approach [34] is still needed, with the forecast error outlier detection method pointing out potential outliers and a human agent validating them. Thus, the effort would be less costly with the application of the method presented in this chapter. This technique has also proven useful in the completeness assessment of the GreEn-ER data.

In addition, this approach relies on high quality predictions, which may be improved. One way to improve the forecasts is using more features in the training phase. In the case of datasets with similar pattern to the one presented in this thesis, it is common to also use the consumption from one week earlier. However, in the present dataset, with several subsequent samples with data quality problems, the use of the past consumption could degrade the model. On the other hand, in a real-time application, this feature could be of great help in the definition of a good predictor and would significantly improve the outlier and anomaly detection.

## References

- [1] Réseau de Transport D'électricité, "Bilan Électrique 2018". 2019. Available online: <https://bilan-electrique-2020.rte-france.com/wp-content/uploads/2019/02/BE-PDF-2018v3.pdf>. Accessed on 17 November 2021..
- [2] Zhou, X.; Yang, T.; Liang, L.; Zi, X.; Yan, J.; Pan, D. Anomaly detection method of daily energy consumption patterns for central air conditioning systems. *J. Build. Eng.* 2021, 38, 102179; ISSN 2352-7102. <https://doi.org/10.1016/j.jobe.2021.102179>.
- [3] Gaur, M.; Makonin, S.; Bajic, I.V.; Majumdar, A. Performance Evaluation of Techniques for Identifying Abnormal Energy Consumption in Buildings. *IEEE Access* 2019, 7, 62721–62733. <https://doi.org/10.1109/access.2019.2915641>.
- [4] Himeur, Y.; Ghanem, K.; Alsalemi, A.; Bensaali, F.; Amira, A. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Appl. Energy* 2021, 287, 116601, ISSN 0306-2619. <https://doi.org/10.1016/j.apenergy.2021.116601>.
- [5] Lee, D.; Lai, C.; Liao, K.; Chang, J. Artificial intelligence assisted false alarm detection and diagnosis system development for reducing maintenance cost of chillers at the data centre. *J. Build. Eng.* 2021, 36, 102110, ISSN 2352-7102, <https://doi.org/10.1016/j.jobe.2020.102110>.
- [6] Lehmann, R.  $3\sigma$ -Rule for Outlier Detection from the Viewpoint of Geodetic Adjustment. *J. Surv. Eng.* 2013, 139, 157–165, [https://doi.org/10.1061/\(asce\)su.1943-5428.0000112](https://doi.org/10.1061/(asce)su.1943-5428.0000112).
- [7] Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley Publishing Company: Boston, MA, USA, 1977; p. 988, ISBN 0201076160.
- [8] Wang, H.; Bah, M.J.; Hammad, M. Progress in Outlier Detection Techniques: A Survey. *IEEE Access* 2019, 7, 107964-108000, doi:10.1109/ACCESS.2019.2932769.
- [9] Zhang, J. Advancements of Outlier Detection: A Survey. *ICST Trans. Scalable Inf. Syst.* 2013, 13, 1–26, <https://doi.org/10.4108/trans.sis.2013.01-03.e2>.
- [10] Satman, M.H. A New Algorithm for Detecting Outliers in Linear Regression. *Int. J. Stat. Probab.* 2013, 2, 101–109. <https://doi.org/10.5539/ijsp.v2n3p101>.
- [11] Abraham, B.; Chuang, A. Outlier Detection and Time Series Modeling. *Technometrics* 1989, 31, 241–248, doi:10.2307/1268821.
- [12] Ho, T.K. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282, doi:10.1109/ICDAR.1995.598994.

- [13] Hubert, M.; Vandervieren, E. An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.* 2008, 52, 5186–5201. <https://doi.org/10.1016/j.csda.2007.11.008>.
- [14] Hawkins, D.M. *Identification of Outliers; Monographs on Applied Probability and Statistics; Springer Netherlands: Dordrecht, The Netherlands, 1980; 188p.* ISBN 9789401539944. <http://dx.doi.org/10.1007/978-94-015-3994-4>.
- [15] Vandepu, N. *Data Science for Supply Chain Forecast; Independently published: Chicago, IL, USA. 2018, ISBN 978-1730969430.*
- [16] Breunig, M.M.; Kriegel, H.-P.; Ng, T.R.; Sander, J. LOF: Identifying Density-based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD, Dallas, TX, USA, 15–18 May, 2000; pp. 93–104, doi:10.1145/335191.335388. ISBN 1-58113-217-4.*
- [17] Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* 2013, 49, 764–766, <https://doi.org/10.1016/j.jesp.2013.03.013>.
- [18] Kimber, A.C. Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions. *J. R. Stat. Soc. Ser. C* 1990, 39, 21–30. <https://doi.org/10.2307/2347808>.
- [19] Straker, D. Measuring Spread. Available online: [http://www.syque.com/quality\\_tools/toolbook/Variation/measuring\\_spread.htm](http://www.syque.com/quality_tools/toolbook/Variation/measuring_spread.htm). Accessed on 17 November 2021..
- [20] Huber, P.J. *Robust Statistics, 1st ed.; Wiley Series in Probability and Statistics; Wiley-Interscience: Hoboken, NJ, USA, 1981.*
- [21] Wickham, H.; Stryjewski, L. 40 Years of Boxplots, *Had.Co.Nz.* 2012. Available online: <https://vita.had.co.nz/papers/boxplots.html>. Accessed on 17 November 2021..
- [22] Olano, X.; de Jalón, A.G.; Pérez, D.; Barberena, J.G.; López, J.; Gastón, M. Outcomes and features of the inspection of receiver tubes (ITR) system for improved O&M in parabolic trough plants. *AIP Conf. Proc.* 2018, 2033, 030011. <https://doi.org/10.1063/1.5067027>.
- [23] Brys, G.; Hubert, M.; Rousseeuw, P.J. A robustification of independent component analysis. *J. Chemom.* 2005, 19, 364–375. <https://doi.org/10.1002/cem.940>.
- [24] Kent, A.; Berry, M.M.; Luehrs, F.U.; Perry, J.W. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *Am. Doc.* 1955, 6, 93–101.
- [25] Van Rijsbergen, C.J. *Information Retrieval, 2nd ed.; Butterworth-Heinemann: Waltham, MA, USA, 1979.*



- [26] Zhao, H.; Tang, Z. The review of demand side management and load forecasting in smart grid. In Proceedings of the 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China, 12–15 June 2016; pp. 625–629, doi:10.1109/WCICA.2016.7578513.
- [27] Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. Classification and Regression Trees; Wadsworth and Brooks: Monterey, CA, USA, 1984.
- [28] HARP. “Harp Random Forests”. Available online: <https://dsc-spidal.github.io/harp/docs/examples/rf/>. Accessed on 17 November 2021..
- [29] Sammut, C.; Webb, G.I. “Mean Absolute Error”. In Encyclopedia of Machine Learning; Springer: Boston, MA, USA, Available online: [https://doi.org/10.1007/978-0-387-30164-8\\_525](https://doi.org/10.1007/978-0-387-30164-8_525). Accessed on 17 November 2021.
- [30] De Myttenaere, A.; Golden, B.; Le Grand, B.; Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing* 2016, 192, 38–48, <https://doi.org/10.1016/j.neucom.2015.12.114>.
- [31] Martin Nascimento, G.F.; Delinchant, B.; Wurtz, F.; Kuo-Peng, P.; Jhoe Batistela, N. “Power Consumption Data Quality”, Available online: <https://gricad-gitlab.univ-grenoble-alpes.fr/martgust/power-consumption-data-quality>. Accessed on 17 November 2021.
- [32] [dataset] Martin Nascimento, G.F.; Delinchant, B.; Wurtz, F.; Kuo-Peng, P.; Jhoe Batistela, N.; Laranjeira, T. GreEn-ER—Electricity Consumption Data of a Tertiary Building. Mendeley Data, V1, 2020. <http://dx.doi.org/10.17632/h8mmnthn5w.1>. Accessed on: 9 December 2021.
- [33] Delinchant, B.; Wurtz, F.; Ploix, S.; Schanen, J.; Marechal, Y. GreEn-ER living lab: A green building with energy aware occupants. In Proceedings of the 5th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS), Rome, Italy, 23–25 April 2016, pp. 1-8.
- [34] Wurtz, F.; Delinchant, B. “Smart buildings” integrated in “smart grids”: A key challenge for the energy transition by using physical models and optimization with a “human-in-the-loop” approach. *Comptes Rendus Phys.* 2017, 18, 428–444, ISSN 1631-0705, <https://doi.org/10.1016/j.crhy.2017.09.007>.

## 5 Non-Intrusive Load Monitoring – State of Art

Chapter 5 presents non-intrusive load monitoring (NILM) techniques from the incipient work of Hart to some others widespread algorithms, such as the Factorial Hidden Markov Model (FHMM) and three others based on Artificial Neural Networks (ANN). These methods will be later used to disaggregate some loads from the GreEn-ER building. This chapter also presents the NILMTK, an open source framework dedicated to promote reproducibility in the NILM field by enabling the performance comparison of several algorithms applied to some of the most popular datasets.

### Chapter Contents

5.1	Direct and Indirect Feedbacks Towards Energy Consumption Savings-----	106
5.2	NILM Methods Based On the Activation of Finite States -----	109
5.3	NILM Methods Based on Artificial Neural Networks -----	112
5.4	The NILM Toolkit – NILMTK-----	121
5.5	Conclusions-----	124

## 5.1 Direct and Indirect Feedbacks Towards Energy Consumption Savings

Energy consumption in buildings is often directly related to inhabitants' behavior. However, it is not easy to the average person to quantify the impact of his behavior on the energy consumption. Hence, people need to rely on feedbacks to understand their consumption and adjust their behavior to achieve energy and financial savings. There are several types of feedback that are sent to people regarding their energy consumption, while the Standard Billing the most common of them. Studies have shown that the savings amount achievement is directly related to the type of feedback received by consumers [1] [2]. The feedbacks can be classified into indirect and direct feedbacks [3]. Indirect feedbacks are information provided after the occurrence of consumption, while the direct ones are provided in real-time (or nearly real-time) [4].

Some examples of indirect feedbacks are:

- **Standard Billing:** The most usual one. Bill received periodically, usually monthly, which displays only the consumption, the charges and the amount due.
- **Enhanced Billing:** Besides the information provided in the standard one, this approach delivers more detailed data, such as historical monthly consumption, for instance.
- **Estimated Feedback:** Uses statistical techniques based on the customer's household type, appliance information and billing data to disaggregate the electricity consumption. It estimates the consumption of major appliances, delivering "per appliance" information.
- **Daily or Weekly Feedback:** Information available more frequently than the standard and enhanced bills, this type of feedback helps identifying overconsumption and efficiency opportunities earlier, increasing potential savings.

On the other hand, examples of direct feedbacks are:

- **Real-Time feedback:** Built-in displays in homes providing aggregated real-time energy consumption and cost.
- **Real-Time Plus Feedback:** Built-in displays showing disaggregated energy consumption and cost in real-time on the appliance level.

Some studies [1] performed in the residential sector estimated the energy saving potential, illustrated in Figure 57, of each of these types of feedback, when compared to the standard billing one. It can be seen in this figure that the more frequent the feedback, the higher the energy saving potential. This is particularly true when considering direct feedbacks, provided in real-time, which is one reason why the large-scale installation of smart meters in buildings is encouraged by several governments.

However, normally these meters only measure the overall consumption of the household, not exploring the full potential of the direct feedbacks. This kind of feedback associated to more detailed ones, such as the “Real-Time Plus” that provides data to the appliance level, allows for the largest savings.

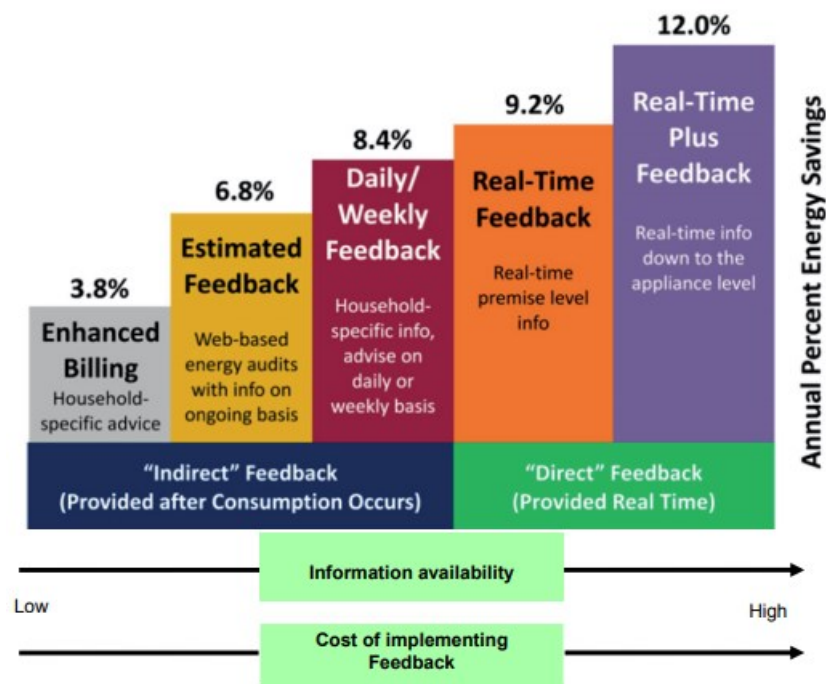


Figure 57 – Energy saving potential of different types of feedback, figure reproduced from [1].

The final section of Chapter 2 showed the prospection of energy sobriety opportunities by analyzing the historical data from individual appliance consumption. This is possible, in the context of the GreEn-ER building, thanks to the historical data available and the massive monitoring down to the appliance level. Nevertheless, this level of detail is no the standard found in most buildings.

The individual appliance consumption can be obtained through comprehensive monitoring, when most loads are measured separately, or through machine learning methods

that use energy disaggregation techniques. The latter type of monitoring is known as NILM (Non-intrusive Load Monitoring). The initial NILM approach was proposed by Hart in the early 1990s in his work entitled “Nonintrusive appliance load monitoring” [5], regarding especially the residential sector. It can be summarized as inferring individual appliances’ energy consumption from the global load curve, obtained at a single measurement point.

With the popularization of smart-meters and the possibility of having a more precise assessment of the electricity consumption of a specific consumer (household, building, etc.), even remotely, the NILM has caught the attention of researchers worldwide. Several datasets were released with the aim to benchmark the algorithms developed to the energy disaggregation task. Some examples were already listed in Chapter 2. Furthermore, most of the public datasets released concern the residential sector, which has allowed for the development of algorithms that are more adapted to the residential environment than to the tertiary sector.

The advance of NILM research in the residential sector relies on four main assumptions. Firstly, at such environment, because of the limited number of appliances, it is rare that two or more appliances change their functioning state simultaneously given an appropriate sampling rate. Because of that feature, many algorithms were based on event (switch of state, for instance) detection [6] [7]. In addition, the power of many residential loads is constant while remaining at the same state. This transforms the load curve into step-wise shape, making it easier to identify the loads upon detection of an event [8]. Furthermore, residential loads often follow a daily and weekly pattern, with respect to the household’s occupants. This periodic behavior has been used to improve the accuracy of the algorithms. Nevertheless, as each household has its own individual schedule, the algorithms need to learn the schedule of all households separately, toughening the transfer learning from one household to another. At last, in addition to temporal features, some loads present strong correlations regarding their usages, which can also be features that are used to help identifying loads from each detected event [9]. One example of these strong correlations is the usage of personal computers and screen monitors that are usually used together.

Several NILM approaches are based on supervised machine learning techniques, which used well label data for training the algorithms. Hence the importance of the datasets presented in Chapter 2.

Therefore, the next section exposes in further details two of the most widespread methods based on event detection: the Combinatorial Optimization, initially proposed by Hart, and the Factorial Hidden Markov Model. Afterwards, three methods based on Artificial Neural Networks are briefly presented. These techniques will be used in the next chapter to disaggregate target loads from the GreEn-ER building's global consumption.

## 5.2 NILM Methods Based On the Activation of Finite States

Assuming that the appliances present in a household respect the assumptions stated earlier (only one appliance at a time change its state and the power on every state is constant), it is sufficient to determine when every appliance works in each state to estimate the consumption profile of the loads from the main load curve. This section addresses two of the most used methods that are based on the activation of finite states: the Combinatorial Optimization, initially proposed by Hart and the Factorial Hidden Markov Model.

### 5.2.1 Combinatorial Optimization

In his incipient work [5], Hart proposed a method, known nowadays as Combinatorial Optimization, which consists in finding the combination of the consumption of all appliances at a given time  $t$  that minimizes the difference between this combination and the overall ground truth consumption. This method was idealized considering that the appliances operate with finite states, with constant consumption in each one of them.

Originally, Hart considered just appliances with two states, on and off. Therefore, considering  $n$  appliances, a vector  $a(t)$  can be defined at each time  $t$  as follows:

$$a_i(t) = \begin{cases} 1, & \text{if appliance } i \text{ is on at time } t \\ 0, & \text{if appliance } i \text{ is off at time } t \end{cases} \quad (1)$$

in which  $i$  represents the indexes of the appliances, going from 1 to  $n$ .

Therefore, the overall power consumption at a given time  $t$  can be modeled as:

$$P(t) = \sum_{i=1}^n a_i(t)P_i + e(t) \quad (2)$$

In which  $P_i$  is the power of the appliance  $i$  in on state and  $e(t)$  is a term to model unaccounted appliances and noise in the measure. If the power consumption of every appliance is known, the disaggregation can be defined as an optimization problem. In this problem a state vector  $A^*$ , that contains the state of each appliance at every timestamp can be estimated by:

$$A^* = \min \left| P(t) - \sum_{i=1}^n a_i(t)P_i \right| \quad (3)$$

This approach was the first method proposed to disaggregate appliances' individual consumption from the global load. However, the popularization of more complex loads (like modern electronic devices) reduced the performance of this technique.

### 5.2.2 Factorial Hidden Markov Model (FHMM)

Another popular algorithm to tackle the disaggregation problem is the Factorial Hidden Markov model. It is based in its core on Markov chains, which are stochastic processes with discrete or continuous state space that presents the Markovian property. This property states that in regular and discrete time intervals, this stochastic process evolves from one condition to another depending only on its last condition, independently of the others.

The Hidden Markov model comes as an extension of the Markov chain. This model includes the case in which the observation is a probability function of the state, i.e., the resulting model is a double-layer stochastic process, in which one stochastic process is underlying and unobservable (hidden) that can only be observed by the other stochastic process that produces the sequences of observations. Due to its flexibility and to the simplicity and efficiency of its parameter estimation algorithm, the hidden Markov model (HMM) has emerged as one of the basic statistical tools for modeling discrete time series, finding widespread application in the areas of speech recognition [11] and molecular computational analyses [12].

The disaggregation problem tackled by the FHMM in this work is to infer:

- $Q^{(1)} = \{q_1^{(1)}, q_2^{(1)}, \dots, q_T^{(1)}\}$
  - $Q^{(2)} = \{q_1^{(2)}, q_2^{(2)}, \dots, q_T^{(2)}\}$
  - $\vdots$
  - $Q^{(M)} = \{q_1^{(M)}, q_2^{(M)}, \dots, q_T^{(M)}\}$
- as the power load of each of the M appliances,
- Such that  $y_t = \sum_{i=1}^M q_t$ ,

- Given  $Y = \{y_1, y_2, \dots, y_T\}$  as the aggregated power for T time periods [22] as the sequence of observations.

A HMM is essentially a mixture model, encoding information about the history of a time series in the value of a single multinomial variable—the hidden state—which can take on one of K discrete values. Figure 58 presents an illustration of the HMM. The system is characterized by this internal discrete state variable, which evolves as a Markov chain between time points. Formally, an HMM can be defined by:

- The finite set of hidden states  $S = \{S_1, S_2, \dots, S_K\}$ ,
- The matrix representing the probability of transitioning from one state to another  $A = \{a_{ij}, 1 \leq i, j \leq N\}$ , being  $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$  with  $a_{ij} \geq 0$  and  $\sum_{j=1}^M a_{ij} = 1$ ,
- The initial state probability distribution  $\pi = \{\pi_i\}$ , being  $\pi_i = P(q_1 = S_i)$

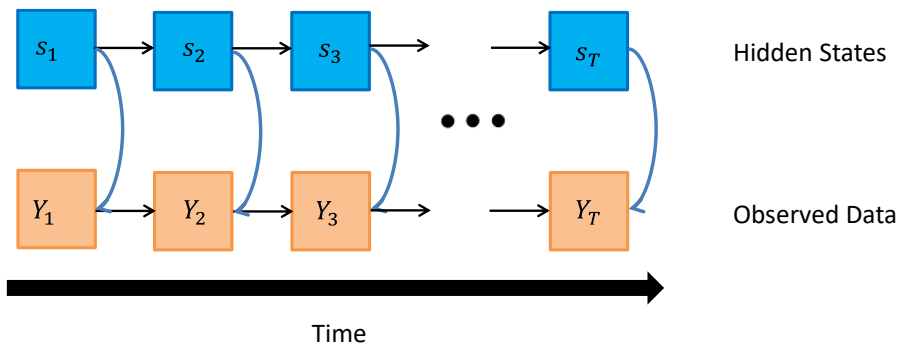


Figure 58 – Illustration of the Hidden Markov Model. Based on Bonfigli and Squartini. [21]

In a standard HMM, the system is characterized by an internal discrete state variable, which evolves as a Markov chain between time points. An extension of the HMM is the Factorial Hidden Markov Model (FHMM) [13]. The FHMM extends the HMM by representing the hidden state in a factored form. This way, the information from the past is propagated in a distributed manner through a set of parallel Markov chains. The parallel chains can be viewed as latent features, which evolve over time according to Markov dynamics.

The FHMM applied to solve the NILM problem [6] has been proven to have a good effect in the disaggregation of residential load with low sampling rate such as 1/60Hz. It does not directly output the observations of each hidden Markov chain, but outputs the sum of the



observations of them. For the NILM problem, the total active power or reactive power is the observation sequence, and the state and power consumption of each appliance are unknown. Therefore, each equipment can be described as a HMM, and the working state of an equipment is a Markov chain. The total power is the sum of the power of all appliances, so it can be described as a FHMM composed of multiple HMMs, and the observation sequence of FHMM is the power consumption. The concept of FHMM to address the NILM problem is illustrated in the Figure 59.

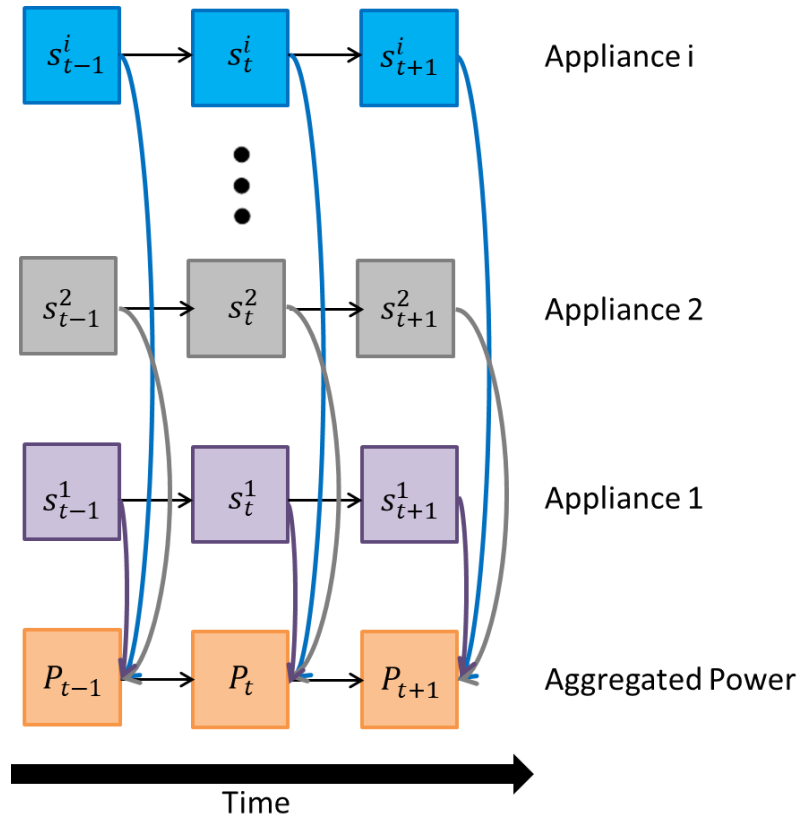


Figure 59 – Illustration of the Factorial Hidden Markov Model. Based on Bonfigli and Squartini. [21]

### 5.3 NILM Methods Based on Artificial Neural Networks

The limitations of the methods presented in the previous sections led the researchers to seek alternatives to tackle the energy disaggregation problem. Allied to this, the advance in computational power allowed the use of more advanced techniques to tackle the issue. One example of that is the development of algorithms based on artificial neural networks (ANN), used with success in other machine learning fields, such as image and language processing, led researchers to start looking for ways to adapt these techniques also to solve NILM problems. This section presents a brief introduction to artificial neural networks and to the most popular layers that can be used to form an artificial neural network, such as fully

connected, convolutional and recurrent neural networks. Afterwards, some NILM algorithms developed based on ANNs are equally presented.

### 5.3.1 Brief Introduction to Artificial Neural Networks

An ANN is a directed graph where the nodes are artificial neurons and the edges allow information from one neuron to another one (or the same neuron in a future time step). An artificial neuron is a model representation that simulates the functioning of a biological neuron in a simplified way. In simple terms, an artificial neuron computes the weighted sum of several inputs, applies an activation function, and passes the result on. These inputs can also be the output from another artificial neurons present in a network. The diagram in Figure 60 illustrates an artificial neuron. In this diagram  $x_n$  represents the inputs, or the data feeding the neuron. Among several inputs, some will stimulate the receiving neuron more and some less. This behavior is simulated by the weights  $w_{kn}$ , which are multiplied by the inputs. An additive junction, with a bias, then sums the weighted inputs. Afterwards, the outcome passes by an activation function, which will define the neuron output. The learning of an ANN happens through the modification of the weights. [14]

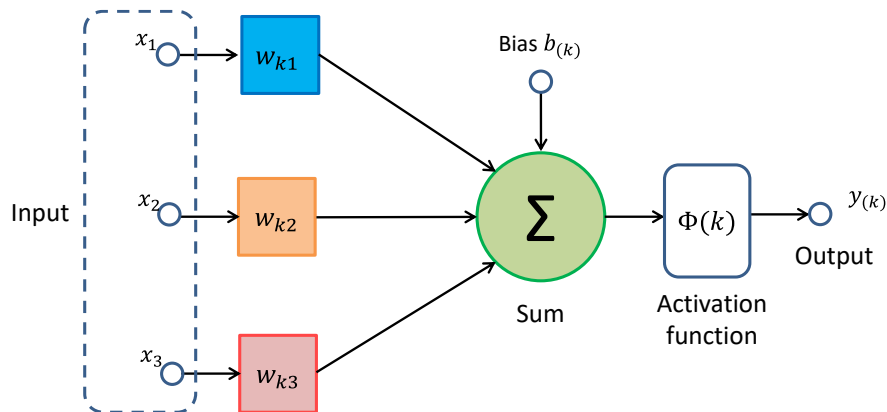


Figure 60 – Representation of an Artificial Neuron. Based on [14]

In an ANN, neurons are typically arranged into layers. An ANN has at least two of them: an input and an output layer. Any layer in between is called a hidden layer. These neural networks can be also classified in shallow and deep networks. In the shallow one, usually only one hidden layer is present. In opposition, a deep neural network has several hidden layers, often of various types. Fully connected, convolutional, recurrent, among others, are examples of the types of layers that can be present in an ANN.

The flow of information from the input layers, through the hidden layers to the output layer is called forward pass. After computing a forward pass through the network, resulting in an output for a given input, the results are compared with the targets. The weights are then modified in the direction to reduce the error between the network output and the targets. This step is called backward pass, and it corresponds to the learning phase. Several metrics can be used to quantify the error between the output and the target, such as the Mean Absolute Error (MAE), the Mean Squared Error (MSE) etc.

### 5.3.1.1 *Feed Forward Fully Connected*

A Feed forward fully connected network consists of a series of fully connected, or dense, layers. In a dense layer each neuron receives input from all neurons of its previous layer and is found to be the most commonly used layer in the models. Its graphical representation can be seen in Figure 61.

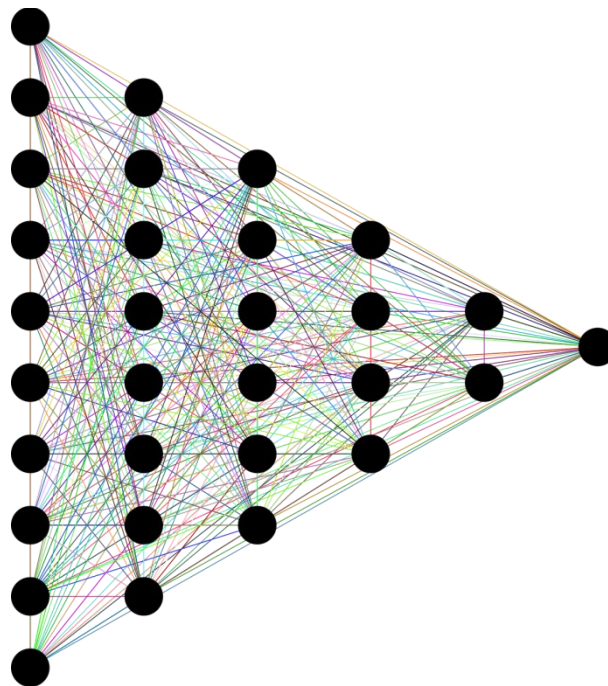


Figure 61 – Representation of a Fully Connected Feed-Forward Artificial Neural Network with three hidden layers.

### 5.3.1.2 *Convolutional layer*

Image recognition is a classic classification problem, and Convolutional Neural Networks have a history of high accuracy for this problem. The biological inspiration for this architecture comes from an experiment performed in 1962 by Hubel and Wiesel [15]. In this experiment, they showed that some neurons are activated together when exposed to some lines or curves, thus producing visual recognition. A Convolutional Network tries to simulate

that by filtering lines, curves, and edges. Each added layer transforms this filtering into a more complex image. The first successful application of a CNN was developed by LeCun in 1998 [16], using seven layers between convolutions and fully connected for online handwriting recognition.

When it comes to image recognition/classification, the inputs are usually two-dimensional (or three if it is a colorful image) matrices with height and width with values for each pixel. The convolutions work like filters that see small squares and "slide" across the image capturing the most striking features. The filter, also known as a kernel, is formed by randomly initialized weights, updating them with each new input during the backpropagation process. Figure 62 shows an illustration of a 2-D convolutional layer.

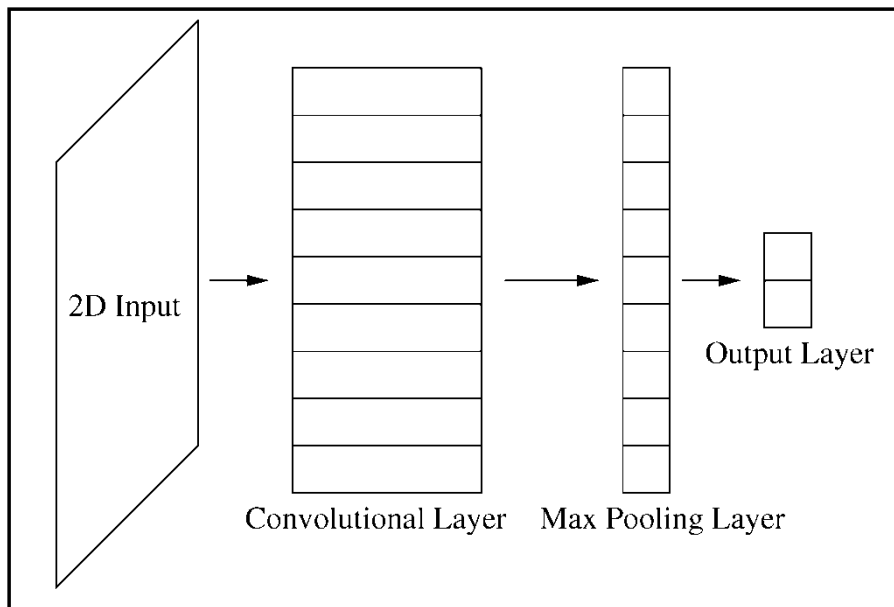


Figure 62 – Representation of a Convolutional Artificial Neural Layer for image processing [23].

Recently some studies demonstrated that 1-D CNNs (1-Dimension Convolutional Neural Networks) are also effective tools for time sequence modelling as it can be treated as a spatial dimension just like the height or width of a two-dimensional image. A schematic illustration can be seen in the Figure 63.

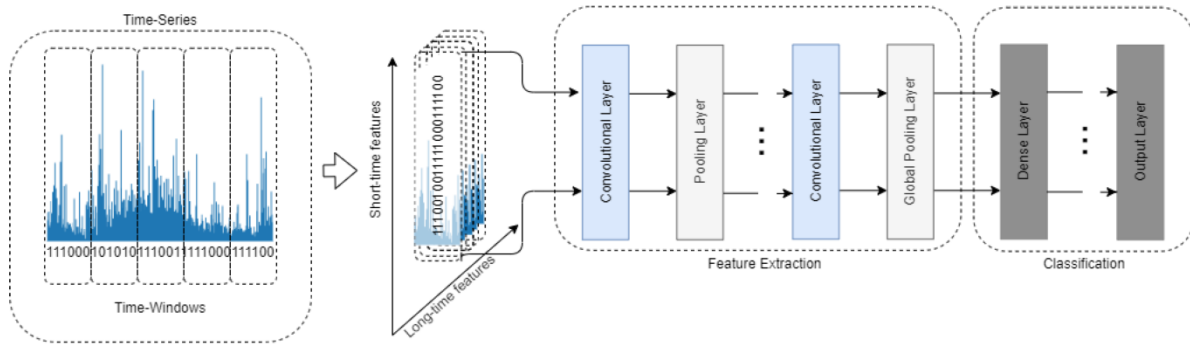


Figure 63 – The conceptual structure of a convolutional neural network used for the purpose of time-series analysis. [27]

It would be possible to learn sequence fragments within a window size series using a convolution window in each convolutional layer to process time series. This process should enable the identification of some subsequences anywhere in the entire time series, so that the local trend change features of the multivariate time series over time can be captured. After the 1D convolution operation, a max pooling operation should be used for subsampling, which outputs the maximum value of subsequences extracted from the input time series. In this way, the length of one-dimensional input time series is reduced.

### 5.3.1.3 Recurrent Neural Network

Feed forward neural networks map from a single input vector to a single output vector. When the network is shown a second input vector, it has no memory of the previous input. Regarding that, the recurrent neural network (RNN) was proposed. A RNN is a type of artificial neural network suited to sequential data or time series data. RNNs support processing of sequential data by the addition of a cycle such that the output from a neuron in a layer  $l$  at a given time step  $t$  is fed via weighted connections to every neuron in layer  $l$  (including neuron itself) at time step  $t + 1$ . In other words, the neurons take as their input not just the current input example they see, but also what they have perceived previously in time. In that manner, they take information from prior inputs to influence the current input and output. This loop allows the network to step through sequential input data whilst persisting the state of nodes in the hidden layer between steps consisting in a sort of working memory. While traditional deep neural networks assume that inputs and outputs are independent from each other, the output of recurrent neural networks depend on the prior elements within the sequence. Another distinguishing characteristic of recurrent networks is that they share parameters across each layer of the network. While feedforward networks have

different weights across each node, recurrent neural networks share the same weight parameter within each layer of the network.

An additional enhancement to RNNs is to use bidirectional layers. In a bidirectional RNN, there are effectively two parallel RNNs, one reads the input sequence forwards and the other reads the input sequence backwards. The output from the forwards and backwards halves of the network are combined by either concatenating them or doing an element-wise sum. Figure 64 illustrates an example of a RNN layer.

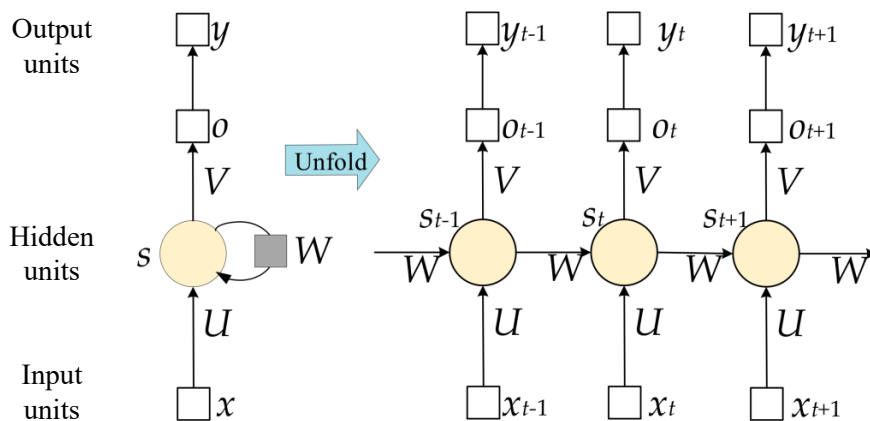


Figure 64 –. Representation of a Recurrent Artificial Neural Layer. [26]

In practice, RNNs may suffer from the ‘vanishing gradient’ problem [17] where gradient information disappears or explodes as it is propagated back through time, which may limit a RNN’s memory.

### 5.3.1.3.1 Long Short-Term Memory Network (LSTM)

Long Short-Term Memory Network or LSTM [17], is a variation of a recurrent neural network (RNN) that is quite effective in predicting the long sequences of data over a period of time. Besides the loop of memory present in the RNNs, it also includes a special unit known as a memory cell to withhold the past information for a longer time for making an effective prediction.

Instead of having a single neural network layer as in a standard RNN, in the LSTM there are four, interacting in a very special way. The key to LSTMs is the cell state. The cell state is like a conveyor belt. It runs straight down the entire chain, with only some minor linear interactions. It’s very easy for information to just flow along it unchanged. The LSTM

does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation. The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through. The first step in our LSTM is to decide what information we’re going to throw away from the cell state. This decision is made by a sigmoid layer called the “forget gate”, represented by  $f_t$  in the diagram. The next step is to decide what new information will be stored in the cell state. This has two parts. First, a sigmoid layer called the “input gate” decides which values will be updated, represented by  $i_t$ . Next, a hyperbolic tangent layer creates a vector of new candidate values,  $\tilde{C}_t$  that could be added to the state. In the next step, these two will be combined to create an update to the cell state.

The output will then be a combination between the input ( $x_t$ ), the output of the previous time step ( $h_{t-1}$ ), and the updated cell state ( $C_t$ ). Figure 65 shows a representation of a LSTM cell.

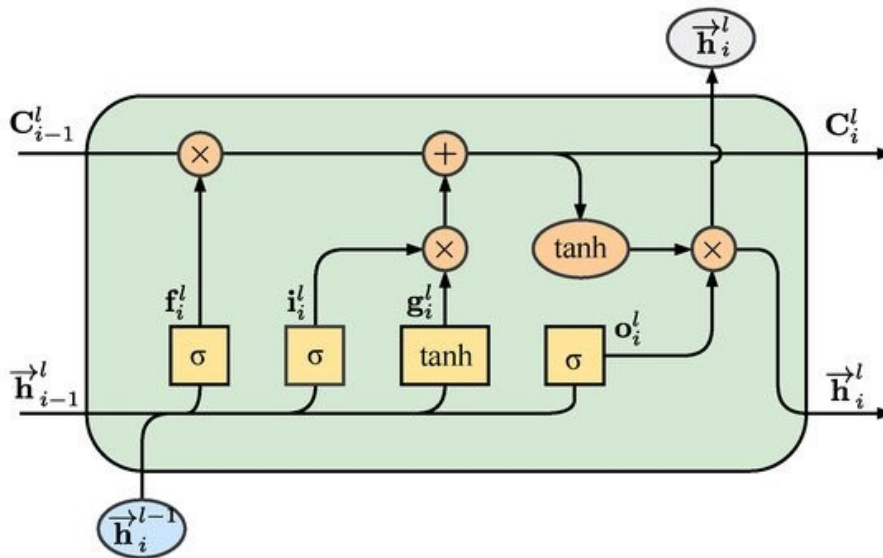


Figure 65 – Representation of a LSTM cell. [24].

### 5.3.1.3.2 Gate Recurrent Unit

A variation on the LSTM is the Gated Recurrent Unit (GRU), introduced by Cho, et al. [18]. It combines the forget and input gates into a single “update gate.” It also merges the cell state and hidden state. A GRU has basically two gates, a reset gate  $r$ , and an update gate  $z$ , as it is shown in Figure 66. Intuitively, the reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the previous memory is

stored. If the reset gate is set to be always one and update gate to be always zero the plain RNN model is retrieved. The basic idea of using a gating mechanism to learn long-term dependencies is the same as in a LSTM.

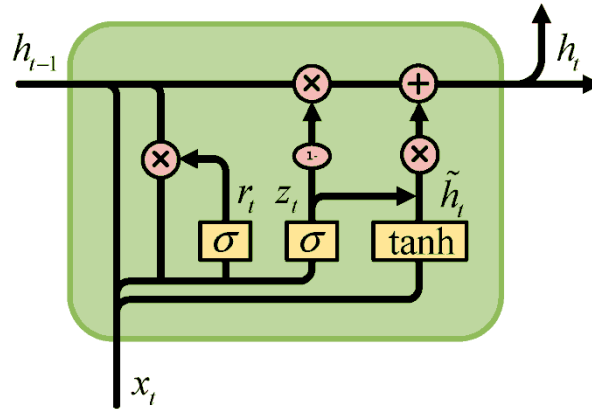


Figure 66 – Representation of a Gate Recurrent Unit cell. [25]

### 5.3.2 NILM algorithms with ANN

In the recent past, most energy disaggregation problems were solved using algorithms based on Hidden Markov Models (HMM), more precisely using their Factorial variation (FHMM), since these models are well suited to sequential data.

However, the popularization of ANN and its success in other machine learning fields such as image and language processing with respect to what had been in use caused researchers to start looking for ways to adapt these techniques also to solve NILM problems. Several algorithms were developed combining different types of layers like RNN with LSTM cells, with GRU cells, or even using convolutional layers. The next sections will detail some of these algorithms.

#### 5.3.2.1 LSTM

An adapted RNN using LSTM layers to perform the energy disaggregation in a residential environment was proposed by Kelly [10]. The architecture of this neural network is shown Figure 67.



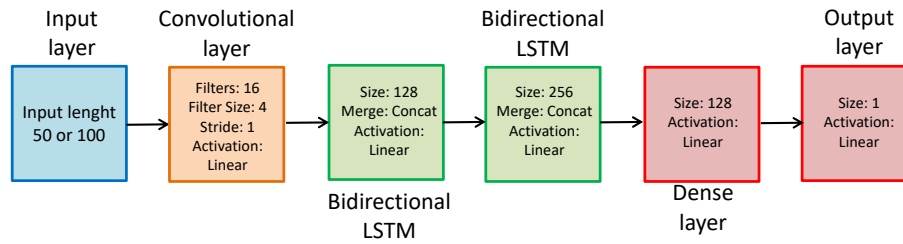


Figure 67 – Artificial Neural Network based on LSTM layers for energy disaggregation. [10]

At each time step, the network sees a single sample of aggregate power data and outputs a single sample of power data for the target appliance. The addition of a convolution layer slightly increase performance (the conv. layer convolves over the time axis). Networks that use LSTM neurons suffer from high computational cost. Each LSTM cell performs several mathematical operations before they produce their output. This makes them computationally demanding for both training and inference. Moreover, they have an internal memory cell which raises the memory demands of the model.

### 5.3.2.2 WindowGRU

The limitations of the LSTM network inspired a new design that performs the same whilst being less demanding. The first step was to replace LSTM neurons with GRU. Gated Recurrent Units have a simpler architecture with no internal memory. This makes them more computationally efficient while training and less memory demanding for disaggregation. Odysseas Krystalakos et al [19] proposed the architecture, shown in Figure 68, to apply in NILM problems.

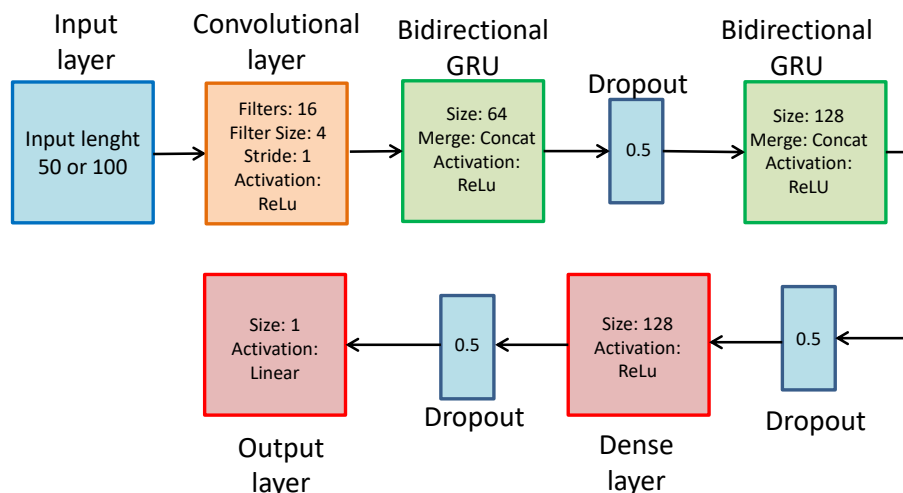


Figure 68 – Artificial Neural Network based on GRU layers for energy disaggregation. [19]

### 5.3.2.3 Sequence-to-point

The neural networks learn a nonlinear regression between a sequence of the mains readings and a sequence of appliance readings with the same time stamps. It can be seen as a sequence-to-sequence approach. These architectures define ANNs that map sliding windows of the input to the corresponding windows of the output target appliance.

Instead of training a network to predict a window of appliance readings, the sequence-to-point [20] method proposes to train a neural network to only predict the midpoint element of a sliding window. The idea is that the input of the network is a sliding window from the aggregated data, and the output is the midpoint element of the corresponding window of the target appliance. This method assumes that the midpoint element is represented as a nonlinear regression of the mains window. The proposed architecture is shown in the Figure 69.

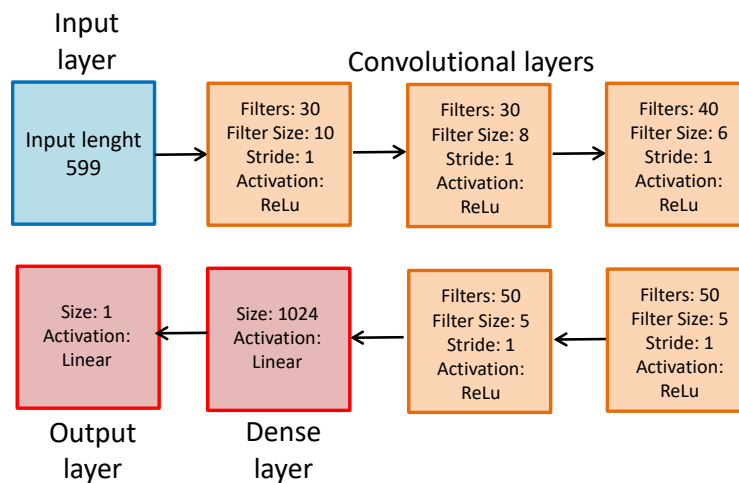


Figure 69 – Sequence-to-point Artificial Neural Network for energy disaggregation. [20]

## 5.4 The NILM Toolkit – NILMTK

The previous sections detailed some of the most popular algorithms used to disaggregate the overall household consumption into individual appliances. From event based algorithms, such as the Combinatorial Optimization and the FHMM, to ANNs approaches, such as the Sequence-to-point. However, the performance comparison between most of the techniques developed is not trivial, since they were developed and tested from different datasets, or different sub datasets from the same dataset, and the evaluation metrics to assess their performance often were not the same.

Because of these issues, the research in the NILM field often lacked of reproducibility. In order to address this concern, Batra et al [28] [29] developed the NILMTK, a framework

designed to enable comparative analysis of different methods using several datasets. In its first version, only the Combinatorial Optimization and the FHMM were implemented. In later versions [30] [31], methods based on ANNs, such as the RNN by Kelly, the Sequence-to-point and the Window GRU, presented in the previous chapter, were also implemented within the framework. Some variations of the FHMM were equally made available among additional variations of other algorithms.

Different datasets usually have their own structure, meaning that the input data format used on the algorithms developed based on them are also different. To address this issue, NILMTK authors proposed a standard data format based on the REDD dataset format, cited in Chapter 2. Parsers dedicated to some of the most popular datasets were implemented to standardize the input data format. Parsers to REDD, UK-DALE, iAWE, COMBED and Dataport datasets, among others, were developed and implemented within the NILMTK.

Then, the parsers convert the data, usually in the CSV format to the Hierarchical Data Format version 5 (HDF5). HDF5, an open source file format, is able to deal with large, complex and heterogeneous data. Additionally, the HDF5 format supports metadata embedding, making it self-descriptive.

The metadata is written in YAML format files, which is a format often used for configuration files. It is through the YAML files that the data is described. These files contain information of the appliances type, the electric schema and the measured quantities: power, energy, voltage, current, etc. There are several YAML files containing metadata. For instance, the NILMTK package for python comes along with another package called `nilm_metadata`. In the later, appliance types are described in YAML files and separated in dedicated files to commercial loads, cooking loads and heating loads, for example.

For instance, there are at least three main “YAML” files containing metadata allowing the parsers to convert the original dataset to a more friendly format to the NILMTK. One file is the “`dataset.yaml`” that describes the dataset, with its name and location among other information. At the same time, “`meter_devices.yaml`” contains the information about the meters. Another file is the “`building1.yaml`”. It is in this file that the loads are described. It contains the information about the electric schema and the type of every load. In the case where there is more than one building, there are other YAML files containing the metadata of each building. These YAML files are exclusive of each dataset. There are, also, other YAML files, containing more general data, such as appliance types.

### 5.4.1 The integration of the GreEn-ER dataset to the NILM-TK

As originally the GreEn-ER dataset was not conceived to integrate the NILM-TK, it was necessary to adapt the dataset to the structure compatible with the NILM-TK. Initially, only the data corresponding to the first layer of the TGBT2 was adapted to the NILM-TK environment.

Every branch in the first layer connected to the TGBT2 was added to the file “commercial.yaml” as an appliance type. They were all described into the “building1.yaml” of the respective parser. Figure 70 presents, as an example, the structure of two YAML files containing metadata of the GreEn-ER dataset for the integration to the NILM-TK. Once the data and the metadata are prepared, the parsers convert these files into a HDF5 file, and the NILM-TK is ready to be used. The usage of the NILM-TK applied to the GreEn-ER data is the object of the next chapter.

<pre>Commercial appliance:   parent: appliance  AHU:   parent: Commercial appliance  elevator:   parent: Commercial appliance  TGBT1:   parent: Commercial appliance  TGBT2:   parent: Commercial appliance  TD-GF:   parent: Commercial appliance  TD2-COM-S:   parent: Commercial appliance  TD2-COM-R:   parent: Commercial appliance  TD2-COM-1:   parent: Commercial appliance  TD2-COM-2:   parent: Commercial appliance  TD2-COM-3:   parent: Commercial appliance</pre>	<pre>instance: 1 original_name: building_1 # GreEn-ER building elec_meters:   1: &amp;Total_building       site_meter: true       device_model: EM6400   2: &amp;TD2-DEM-40       submeter_of: 1       device_model: EM6400   3: &amp;TD2-COM-S       submeter_of: 1       device_model: EM6400   4: &amp;TD2-COM-R       submeter_of: 1       device_model: EM6400  appliances: - original_name: Total_building   type: TGBT2   instance: 1   meters: [1] - original_name: TD2-DEM-40   type: TD2-DEM-40   instance: 1   meters: [2] - original_name: TD2-COM-S   type: TD2-COM-S   instance: 1   meters: [3] - original_name: TD2-COM-R   type: TD2-COM-R   instance: 1   meters: [4]</pre>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

a) Commercial.yaml describing every branch of the TGBT2 as an appliance type

b) Building1.yaml describing the electric schema and the loads

Figure 70 – YAML files containing GreEn-ER dataset metadata for the integration to the NILM-TK.

## 5.5 Conclusions

This chapter presented some of the most popular algorithms developed over the years by several researches in the energy disaggregation domain. Different approaches have been used, such as event detection based algorithms (Combinatorial Optimization and FHMM) and artificial neural networks (LSTM, windowGRU and Sequence-to-point), which were exposed in more details.

Additionally, this chapter addressed the NILMTK, a framework dedicated to the reproducibility of the experiments in the NILM field. It put together several algorithms and datasets under a standard data format in order to enable direct comparison between algorithms applied to different datasets. Hence, several dataset converters are available within the framework in order to format all the input data into a friendly format to the framework. Therefore, to use the GreEn-ER dataset along with the NILMTK, the development of a parser dedicated to this dataset was needed. Metadata was written into YAML files and a converter based on the one available for the REDD dataset was developed. The next chapter addresses the application of the NILMTK to the GreEn-ER dataset.

## References

- [1] K. Ehrhardt-Martinez, K. A. Donnelly and J. A. Laitner. “Advanced metering initiatives and residential feedback programs: A meta-review for household electricity-saving opportunities”. Tech. Rep. E105, American Council for an Energy-Efficient Economy, Washington, DC, 2010.
- [2] Electric Power Research Institute (EPRI). “Residential Electricity Use Feedback: A Research Synthesis and Economic Framework”. Palo Alto, CA: 2009. 1016844.
- [3] S. Darby. “Making it Obvious: Designing Feedback into Energy Consumption”. In: P. Bertoldi, A. Ricci, and A. de Almeida. *Energy Efficiency in Household Appliances and Lighting*. Springer, Berlin, Heidelberg 2001. [https://doi.org/10.1007/978-3-642-56531-1\\_73](https://doi.org/10.1007/978-3-642-56531-1_73)
- [4] G. Wood, and M. Newborough, “Dynamic energy-consumption indicators for domestic appliances: environment, behaviour and design”, *Energy and Buildings*, vol. 35(8), pp 821–841, 2003. [https://doi.org/10.1016/s0378-7788\(02\)00241-4](https://doi.org/10.1016/s0378-7788(02)00241-4)
- [5] G. W. Hart, "Nonintrusive appliance load monitoring," in *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870-1891, Dec. 1992, doi: 10.1109/5.192069.
- [6] J. Z. Kolter and T. Jaakkola. Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1472–1482, La Palma, Canary Islands, 2012
- [7] J. Liang, S. Ng, G. Kendall, and J. Cheng. Load Signature Study - Part I: Basic Concept, Structure, and Methodology. *IEEE Transactions on Power Delivery*, 25(2):551–560, 2010.
- [8] M. J. Johnson and A. S. Willsky. Bayesian Nonparametric Hidden Semi-Markov Models. *Journal of Machine Learning Research*, 14:673–701, 2013.
- [9] H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han. Unsupervised Disaggregation of Low Frequency Power Measurements. In *Proceedings of 11th SIAM International Conference on Data Mining*, pages 747–758, Mesa, AZ, USA, 2011.
- [10] J. Kelly and W. Knottenbelt. Neural NILM: Deep Neural Networks Applied to Energy Disaggregation. In *Proceedings of the 2Nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys '15)*. ACM, New York, NY, USA, 55--64. 2015. <https://doi.org/10.1145/2821650.2821672> event-place: Seoul, South Korea.

- [11] L. Rabiner and B. Juang, "An introduction to hidden Markov models," in *IEEE ASSP Magazine*, vol. 3, no.1, pp. 4-16, Jan 1986, doi: 10.1109/MASSP.1986.1165342.
- [12] Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol.* 1994 Feb 4;235(5):1501-31. doi: 10.1006/jmbi.1994.1104. PMID: 8107089.
- [13] Z. Ghahramani and M. I. Jordan. Factorial Hidden Markov Models. *Machine Learning* **29**, 245–273 (1997). <https://doi.org/10.1023/A:1007425814087>
- [14] McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–133 (1943). <https://doi.org/10.1007/BF02478259>
- [15] D. H. Hubel and T. N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." *The Journal of physiology* vol. 160,1 (1962): 106-54. doi:10.1113/jphysiol.1962.sp006837
- [16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [17] S; Hochreiter and J. Schmidhuber. Long Short-term Memory. *Neural computation*. 9. 1735-80. 1997. doi:10.1162/neco.1997.9.8.1735.
- [18] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. of SSSST*, 2014.
- [19] O. Krystalakos, C. Nalmpantis and D. Vrakas. Sliding Window Approach for Online Energy Disaggregation Using Artificial Neural Networks. 1-6. 2018. doi: 10.1145/3200947.3201011.
- [20] C. Zhang, M. Zhong, Z. Wang, Ni. Goddard, and C. Sutton. Sequence-to-point learning with neural networks for nonintrusive load monitoring. *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, Feb. 2-7, 2018.
- [21] R. Bonfigli and S. Squartini. HMM Based Approach. In *Machine Learning Approaches to Non-intrusive Load Monitoring*, Springer, pp. 31-90, 2020.
- [22] H. Kim, M. Marwah, M. Arlitt, G. Lyon, J. Han. Unsupervised disaggregation of low frequency power measurements. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, Phoenix, AZ, USA, 28–30 April 2011; pp. 747–758.
- [23] G. Bologna. A Simple Convolutional Neural Network with Rule Extraction. *Applied Sciences*. 2019; 9(12):2411. <https://doi.org/10.3390/app9122411>

- [24] N. Minh-Tuan, Y.-H Kim,. Bidirectional Long Short-Term Memory Neural Networks for Linear Sum Assignment Problems. *Appl. Sci.* 2019, 9, 3470. <https://doi.org/10.3390/app9173470>
- [25] F. Pan, J. Li, B. Tan, C. Zeng, X. Jiang, L. Liu, J. Yang. Stacked-GRU Based Power System Transient Stability Assessment Method. *Algorithms* 2018, 11, 121. <https://doi.org/10.3390/a11080121>
- [26] G. Li, H. Wang, S. Zhang, J. Xin, H. Liu. Recurrent Neural Networks Based Photovoltaic Power Forecasting Approach. *Energies* 2019, 12, 2538. <https://doi.org/10.3390/en12132538>
- [27] K. Filus, A. Domański, J. Domańska, D. Marek, J. Szyguła. Long-Range Dependent Traffic Classification with Convolutional Neural Networks Based on Hurst Exponent Analysis. *Entropy* 2020, 22, 1159. <https://doi.org/10.3390/e22101159>
- [28] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring. In: 5th International Conference on Future Energy Systems (ACM e-Energy), Cambridge, UK. 2014. DOI:10.1145/2602044.2602051. arXiv:1404.3878.
- [29] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring". In: NILM Workshop, Austin, US. 2014 [pdf]
- [30] J. Kelly, N. Batra, O.r Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava. Demo Abstract: NILMTK v0.2: A Non-intrusive Load Monitoring Toolkit for Large Scale Data Sets. In the first ACM Workshop On Embedded Systems For Energy-Efficient Buildings, 2014. DOI:10.1145/2674061.2675024. arXiv:1409.5908.
- [31] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson. Towards reproducible state-of-the-art energy disaggregation. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19). Association for Computing Machinery, New York, NY, USA, 193–202. 2019. DOI:10.1145/3360322.3360844





## 6 Disaggregation methods applied to a tertiary building environment – The GreEn-ER case

This chapter starts by discussing the differences between the residential and tertiary environments with respect to the loads present in each field, the challenges met in the disaggregation task in tertiary buildings and some insights to turn around these issues. This chapter presents the results obtained by the application of the NILM techniques presented in the previous chapter to the GreEn-ER dataset. For this, the NILMTK, a framework developed to compare different algorithms and different datasets, was used. Since there are more than 300 meters it is practically impossible to disaggregate all loads into appliance level. Because of that, a subset containing three major consumers from the TGBT2 was chosen as target appliances. The air compressors, the chillers and the restaurant were selected as target loads while the remaining were aggregated into a “Fantôme” or “Ghost” load. The results show that the two techniques based on ANNs presented the best results overall, however at a great computational cost. For simple loads, such as the air compressor handled in this chapter, the results were satisfactory.

### Chapter Contents

6.1	Energy Disaggregation in the Tertiary Sector -----	130
6.2	Energy Disaggregation Applied to GreEn-ER loads -----	132
6.3	Conclusions-----	143

## 6.1 Energy Disaggregation in the Tertiary Sector

Even though the benefits of employing NILM techniques in tertiary and industrial facilities have been acknowledged since the field's inception, most of the algorithms developed were concentrated on residential buildings [3]. They were developed based on residential datasets, being naturally more adapted to tackle the appliances found in residential environments.

Nevertheless, these approaches based on event detection are not well suited to the commercial environment. Reports by the Electric Power Research Institute (ERPI) from the USA [1][2] and the California Energy Commission's Public Interest Energy Research (PIER) [4] program state that these algorithms lacked performance when dealing with multistate and variable-power devices (which can be seen as one extreme form of a multi-state device and violates the assumption of constant power draw of the loads), loads that are largely present in tertiary buildings. Also, commercial buildings have many more appliances than a single household, which means that states' switch are more frequent, making it much more likely to exist more than one appliance changing its state simultaneously. Another typical situation in this type of facility is the presence of several similar loads (such as several personal computers, a typical situation found in office buildings, for instance), which also impairs performance of this type of approach.

In addition, when using high-frequency features to tackle the energy disaggregation, additional harmonic content, which can be introduced by power factor correction devices, micro generation (micro-turbines and photovoltaic panels) and battery storage technologies (electric vehicles included), may also complicate the identification of individual appliances.

Another limitation that can be cited is the building power consumption compared with the power consumption of individual appliances. Small step changes are nearly impossible to identify and segregate from the global consumption. It can be hard to tell if these changes are caused by switching on a desktop computer or if it is due to larger loads changing their demand. These small loads, in which can be included chargers, laptop computers, among other electronic loads, are responsible for significant portion of tertiary building's electricity consumption [5]. Furthermore, the presence of a mix of single-phase and three-phase loads may also harm algorithms' performance, depending on the features used.

It is important to note that several approaches use measurements with relatively high sampling rate, often one second or even sub-second intervals. However, popularizing these

approaches is not trivial, since most facilities have measured data with sampling rate lower than this, such as one, ten and 30 minutes, or even one hour. Because of this, the interest of developing and testing energy disaggregation for low sampled data is even bigger.

Some studies have tried to tackle the energy disaggregation outside the residential environment with low sampled data. The first attempt was applying the same algorithms developed to residential buildings to loads inside the tertiary sector. Batra et al [6] applied the Combinatorial Optimization to disaggregate the consumption of Air Handling Units (AHUs) from the global consumption, and later from an intermediate transformer that is responsible for delivering electricity to the floor where they are located. Their results showed that this algorithm failed to disaggregate the continuously varying demand of the AHUs from the global consumption. However, when the disaggregation is performed at the intermediate transformer level, the results are improved.

Furthermore, Bandeira de Melo Martins developed, in his master's thesis [7], a deep learning algorithm based on convolutional ANNs to disaggregate eight appliances of a factory in Brazil. He successfully disaggregated loads that the FHMM did not perform well, including some exhaust fans, which present a variable-continuous load profile. Nevertheless, his dataset contained a lot of features, such as voltage, current, active power, reactive power, apparent power and active energy consumption, at one second sampling interval. Eight loads were chosen, among which are pelletizers and exhaust fans, because of their share of consumption and importance to the facility. While the FHMM, which is used to compare the results obtained by his method, was able to disaggregate the pelletizers, it failed with the exhaust fans.

In another approach, Ling et al [8] developed a method based on Random Forest algorithm to disaggregate four main loads (lightning, elevators, HVAC and plug-in loads). Despite the promising results, the ground truth used to assess the method were estimated from building occupancy schedule (for lightning, elevators and plug-in loads) and from the cooling and heating loads, simulated by a software and rated performance of appliances (chillers and water pumps). This fact makes these loads have similar profiles, with predictable daily and weekly variations, which is highly unlikely, especially for the elevators.

Summarizing, those three studies were capable to disaggregate the loads from the global consumption under specific circumstances, leading to some insights for the advance of the field. For instance, the Batra study indicated that to tackle the problem of the number of

charges, some level of sub metering could be interesting, leading to a kind of hybrid solution or semi-intrusive load monitoring. In addition, Bandeira de Melo Martins' thesis showed that the choice of target loads is crucial similarly to the results obtained by Ling's approach. In large commercial or industrial buildings, the biggest consumers are more likely to hide the greatest energy saving potential, indicating that these loads are fit candidates to be the target loads in a NILM approach.

## 6.2 Energy Disaggregation Applied to GreEn-ER loads

The previous chapter has presented some of the most popular algorithms used to disaggregate a main consumption into individual appliances. These algorithms were developed regarding residential buildings and their typical behavior. Nevertheless, the load behavior of tertiary buildings is different from residential ones, because of the operating profile (continuously variable or multi-state loads), the number of appliances (many simultaneous state changes) and because of the presence of small loads (compared to the overall consumption). Because of that, it is unsure that the algorithms exposed in Chapter 5 present satisfactory results when applied to tertiary buildings.

Since there are more than 300 hundred metering points inside the GreEn-ER building, it is virtually impossible to disaggregate the global consumption to the appliance level. Several issues play against the successful appliance disaggregation from the global power. The presence of small loads compared to the overall consumption, such as the lightning or the outlets of an individual switchboard, make the disaggregation task more difficult. In addition, there is the issue of poor data quality. Although previous chapters addressed the issue of the data quality, in particular the identification of outliers, it is impractical to pre-treat all measure points, since the algorithm presented in Chapter 4 relies on consumption prediction. Because of the energy meters resolution, 1kWh, the load curve obtained from these meters, in the case of small loads, is not accurate, harming the predictors performance and, hence, the pre-treatment algorithm presented in Chapter 4.

Because of this, a subset of the GreEn-ER dataset was selected to test the NILM techniques embedded into the NILMTK in a tertiary environment. Three first layer of meters connected to the TGBT2 was selected. Eventual outliers from this sub dataset were eliminated using the algorithm presented in Chapter 4.

In order to identify the best approach to perform the energy disaggregation in an environment like the GreEn-ER building, the algorithms presented in the previous chapter were applied with the help of the NILMTK. Methods based on finite states, such as the Combinatorial Optimization and the FHMM, presented in the previous chapter, were compared with some techniques using ANNs, such as the one using the LSTM architecture, the sequence-to-point and the Window GRU. The subdataset used contained one year worth of data, from which nine months were used as training phase and three as test phase. The results were assessed using the Mean Absolute Error (*MAE*), and the Percentage Error (*PE*) between the average values, regarding all the test periods.

In Chapter 5, it was acknowledged that current disaggregation techniques may struggle when addressing tertiary buildings' typical loads. The high amount of loads and continuously variable appliances can be cited as reasons to the poor performance of these NILM algorithms. Nevertheless, some concessions can be made in order to enhance the disaggregation performed by these algorithms in the context of a tertiary building.

One way to improve the results of energy disaggregation algorithms when applied to a tertiary environment is to choose target appliances. By choosing target appliances, the number of loads to be disaggregated decreases and smaller loads combined (those whose power is small compared to overall consumption) become more significant. In addition, considering energy efficiency measures, it is more likely to find potential savings when dealing with major consumers. Therefore, it would be more important to well disaggregate only the major consumers than disaggregate a high amount of loads with average performance.

Hence, this simulation focused in the three biggest consumers linked to the TGBT2. They correspond to the air compressors, the chillers (TD2-GF) and an area called Crous, where the restaurant and the dining hall are located. The rest of consumption linked to this transformer was aggregated and called "fantôme". A diagram of the TGBT2 (switchboard linked to the Transformer 2) highlighting the target loads is shown in Figure 71. At the illustration it can be seen the target loads and the others that form the aggregated load "Fantôme". The consumption profiles of the target loads are also presented in the following figures.

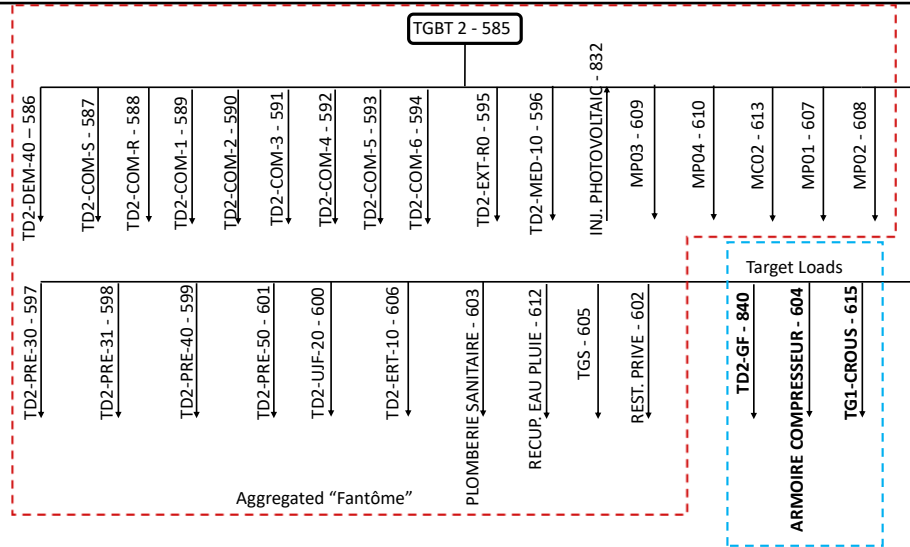
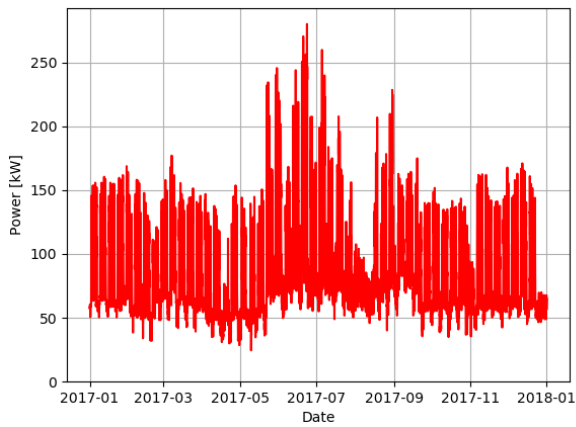
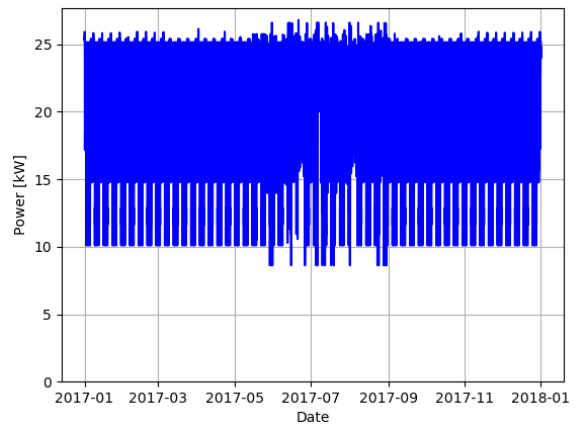


Figure 71 – Electric schema of the TGBT2 with the target loads highlighted.

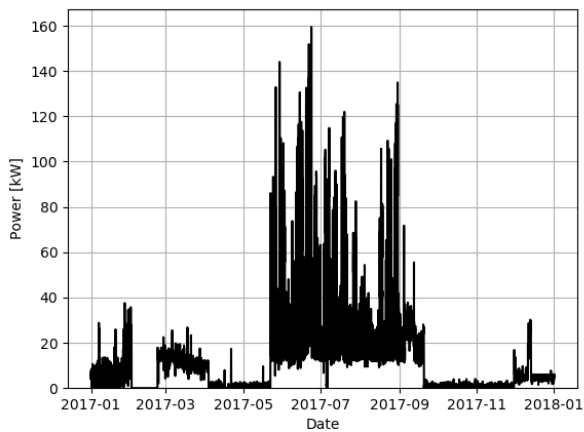


TGBT2

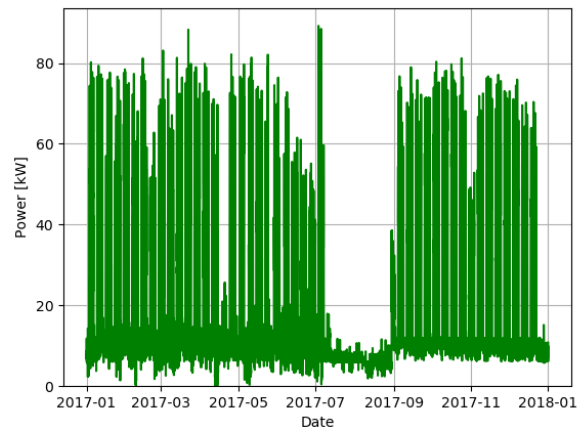


Air compressors

Figure 72 – Load curve of the Air Compressors switchboard and the CROUS switchboard.



TD-GF



CROUS

Figure 73 – Load curve of the TD-GF switchboard and the CROUS switchboard.

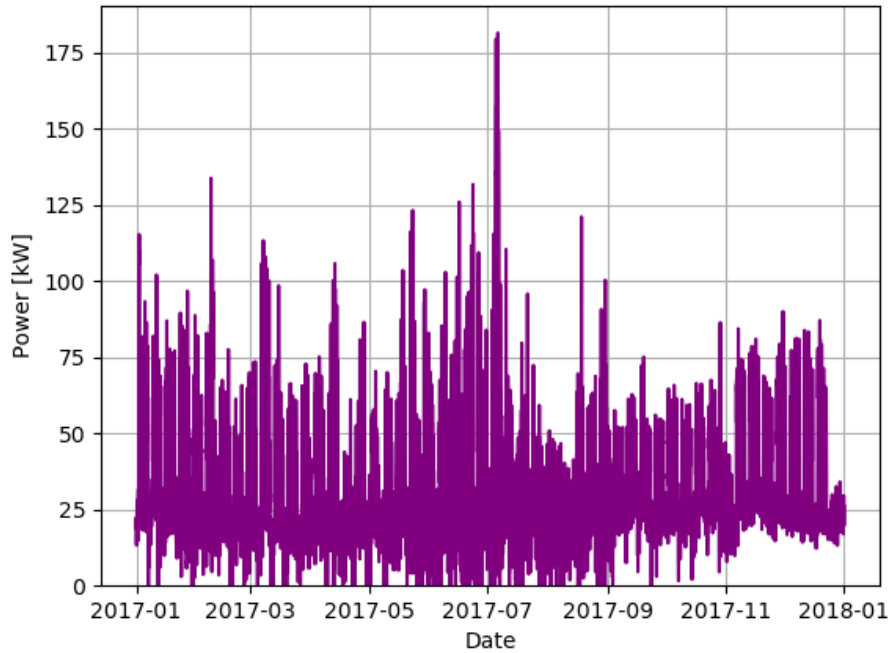


Figure 74 – Load curve of the aggregated load “Fantome”.

Table 24 - Average Power of the target loads.

Load	Average Power [kW]
Air Compressor	21.79
CROUS	17.06
TD-GF	12.98
Fantôme	31.21
<b>TGBT2</b>	<b>83.04</b>

Two air compressors form the compressed air system, one being usually in stand-by while the other is operating. One of them is a fixed-speed rotary screw compressor, which operates usually modulating between two states, while the other one has variable speed, presenting a more continuous curve. During the first half of the year, the fixed-speed one was operating, while the variable speed was in stand-by. This operating mode has been changed along the year, with the variable speed in operation and the other one in stand-by. In order to be fair with the disaggregation algorithms, and not train the algorithm based on one air compressor and test in another completely different, synthetic data based on the fixed-speed compressor replaced the variable-speed data. The global consumption of the TGBT2 was also corrected.

The following sections present the results obtained by performing the disaggregation on these loads applying the cited algorithms.



## 6.2.1 Combinatorial Optimization

The NILMTK was used to apply the combinatorial optimization algorithm to disaggregate the consumption of the target loads from the TGBT2 load curve. The results are presented in the following figures, and in Table 25. In those figures, CO stands for combinatorial optimization while GT stands for ground truth, or the actual value.

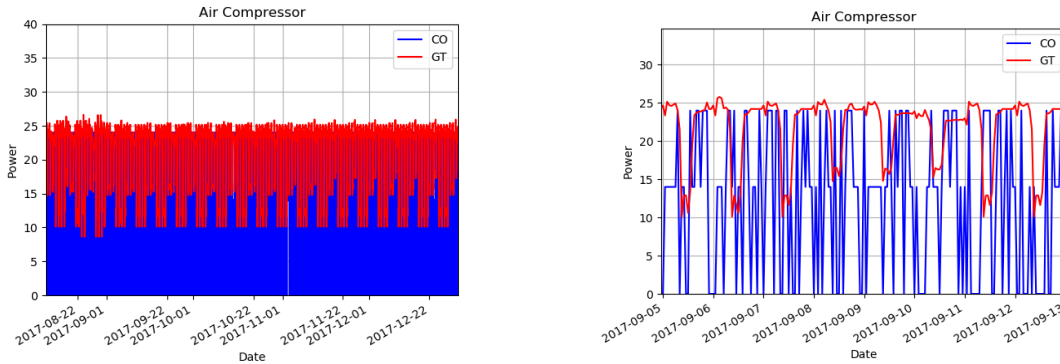


Figure 75 – Disaggregation results of the air compressor using combinatorial optimization.

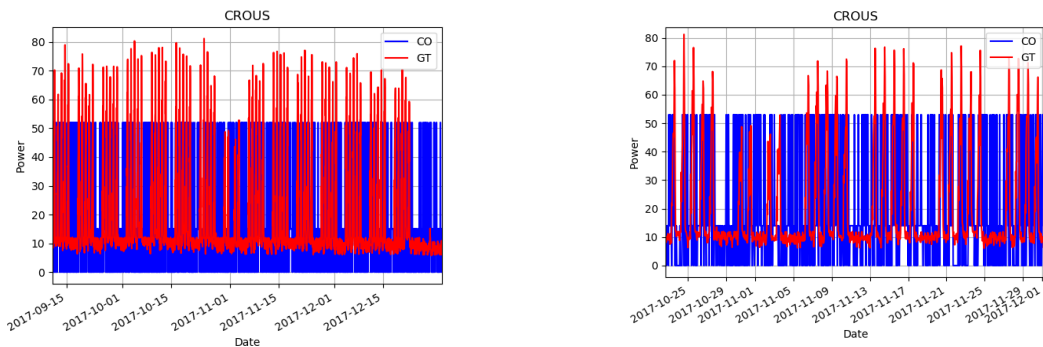


Figure 76 – Disaggregation results of the CROUS using combinatorial optimization.

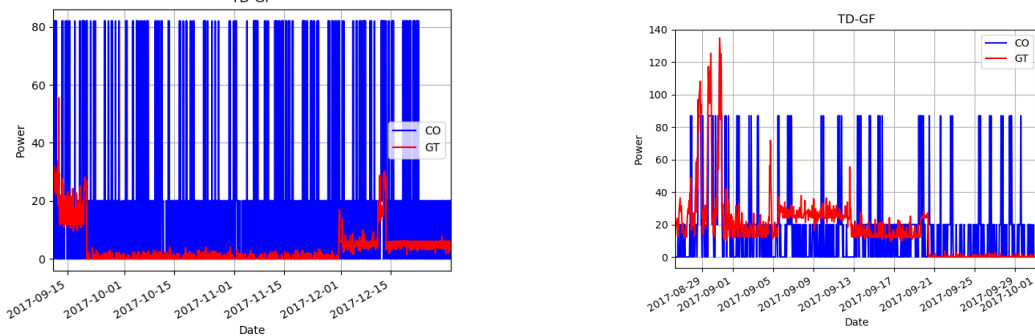


Figure 77 – Disaggregation results of the TD-GF using combinatorial optimization.

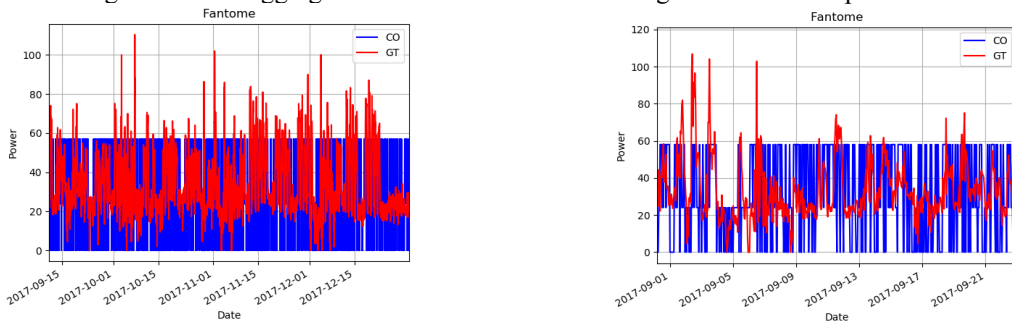


Figure 78 – Disaggregation results of the “fantôme” load using combinatorial optimization.

Table 25 – Combinatorial optimization disaggregation results.

Load	Average power [kW]		MAE [kW]	Percentage error
	Ground Truth	Disaggregated		
Air Compressor	21,80	13,23	10.54	39,31%
CROUS	19,32	19,63	17.12	1,60%
TD-GF	3,44	17,97	17,45	422,38%
Fantôme	33,18	27,44	16,87	17,30%

### 6.2.2 FHMM

The Factorial Hidden Markov Model was also applied to disaggregate the consumption of the target loads. The results are presented in the next figures as well in Table 26.

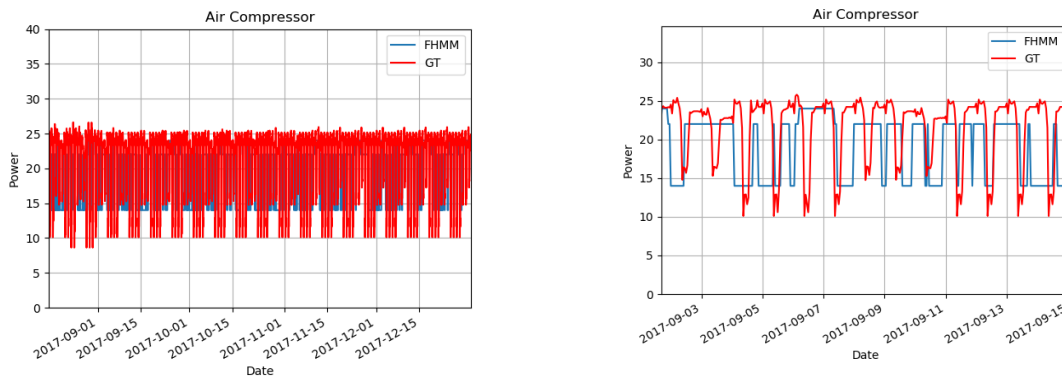


Figure 79 – Disaggregation results of the Air Compressor using FHMM.

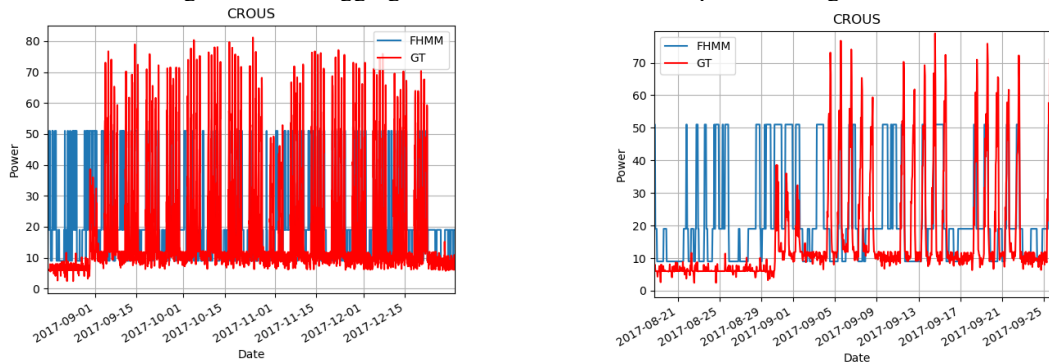


Figure 80 – Disaggregation results of the CROUS using FHMM.

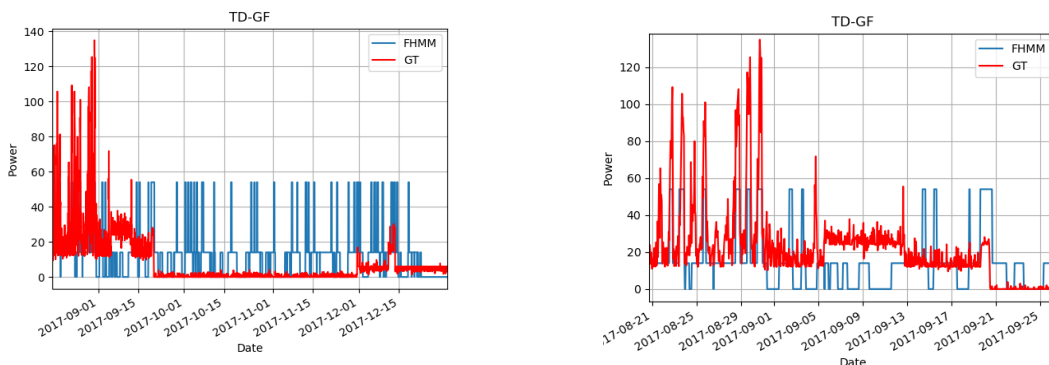


Figure 81 – Disaggregation results of the TD-GF using FHMM.

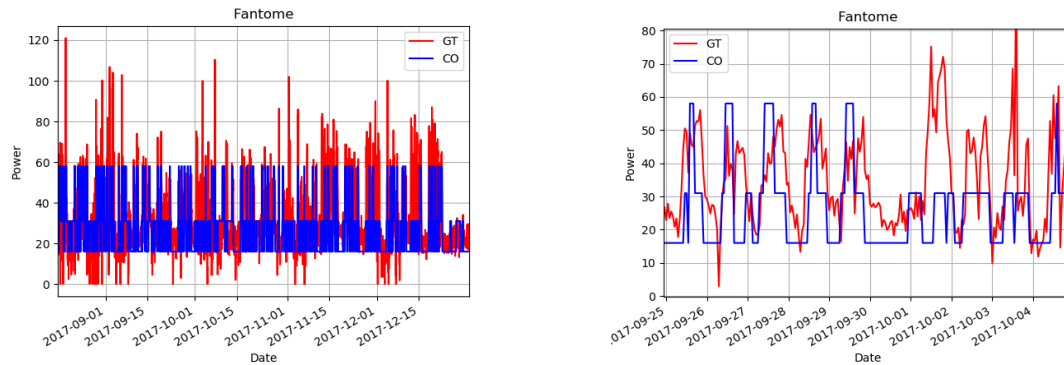


Figure 82 – Disaggregation results of the “Fantôme” load using FHMM.

Table 26 – FHMM disaggregation results.

Load	Average power [kW]		MAE [kW]	Percentage error
	Ground Truth	Disaggregated		
Air Compressor	21,80	18.76	5.88	13,94%
CROUS	19,32	16.99	8.42	12,06%
TD-GF	3,44	14.23	12.59	313,66%
Fantôme	33,18	28.55	11.67	13,95%

The results presented in the section show an improvement when compared to the previous algorithm. It has presented lower MAE for all loads. It has also presented lower percentage error between the average power values for the air compressor. In addition, this technique estimated better the power associated to the states of the Air Compressor than the Combinatorial Optimization. Considering the CROUS load, the percentage error was slightly higher when compared to the previous approach, compensated by the lower MAE.

### 6.2.3 LSTM

The Kelly architecture using the LSTM layers presented in the item 5.3.2.1 was applied to the data. A batch size of 1024 and 50 epochs were used as parameters to configure the training phase. The results obtained are presented in the next figures and in the following table.

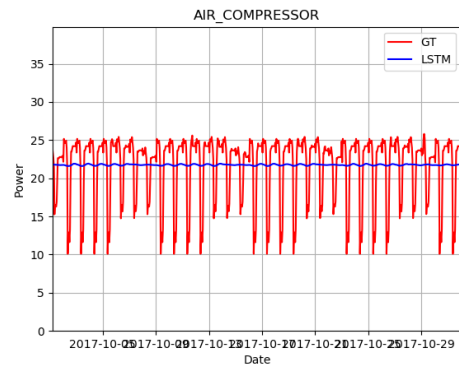
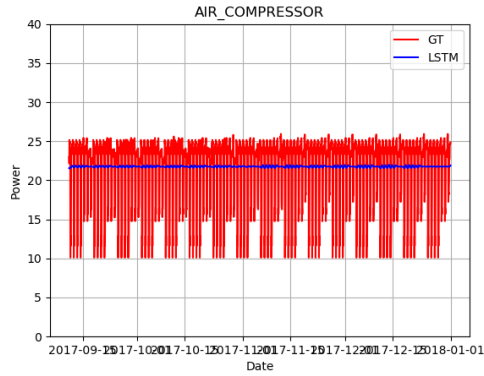


Figure 83 – Disaggregation results of the Air Compressor using LSTM architecture.

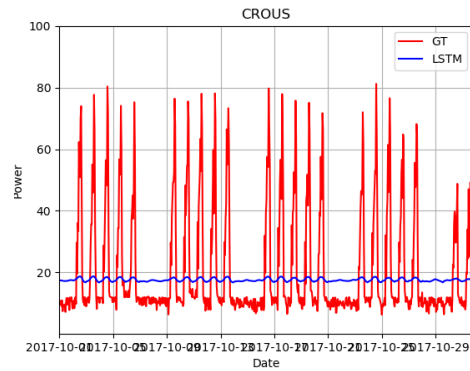
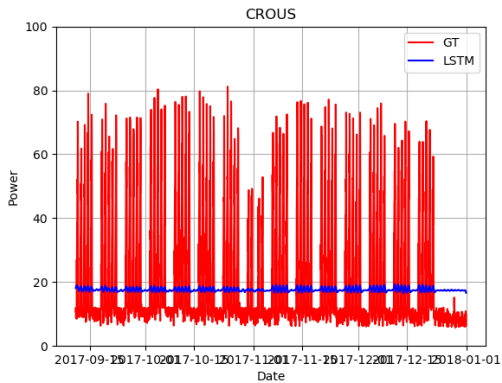


Figure 84 – Disaggregation results of the CROUS using LSTM architecture.

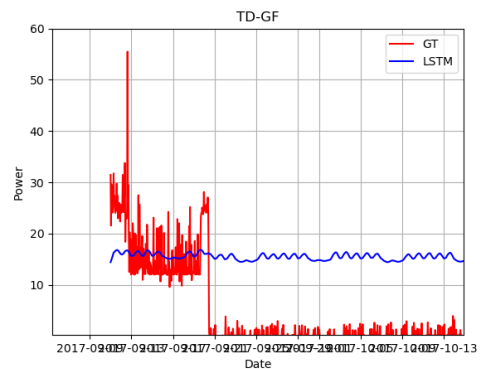
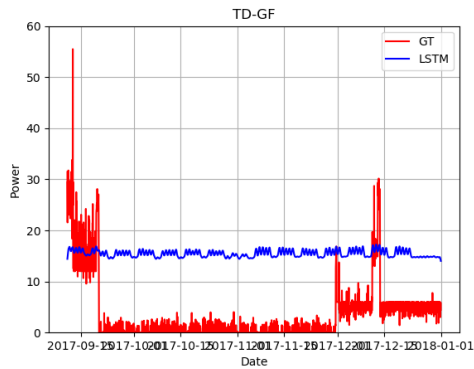


Figure 85 – Disaggregation results of the TD-GF using LSTM architecture.

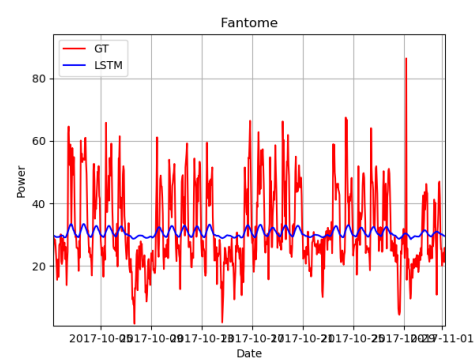
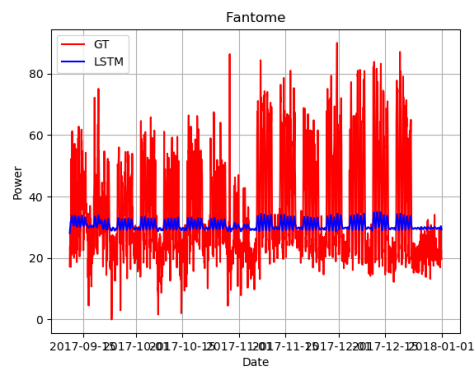


Figure 86 – Disaggregation results of the “Fantôme” load using LSTM architecture.

Table 27 – LSTM-based network disaggregation results.

Load	Average power [kW]		MAE [kW]	Percentage error
	Ground Truth	Disaggregated		

Air Compressor	21,80	21,75	3,40	0,23%
CROUS	19,32	17,64	12,33	8,70%
TD-GF	3,44	15,45	12,72	349,13%
Fantôme	33,18	30,71	11,40	7,44%

Despite presenting good results predicting the energy consumption, translated in the average value and low percentage error, it can be observed in the figures that this algorithm was not able to well reproduce the load curve, underestimating the peaks and overestimating the valleys. However, it can be seen that this algorithm was able to reproduce the seasonal pattern of the data, except for the TD-GF load.

### 6.2.4 Window GRU

The configurations of batch size and the number of epochs for the training phase were maintained as 1024 and 50, respectively, for the Window GRU algorithm, as the same used in the LSTM architecture. The results obtained are exposed in the next figures and in Table 28.

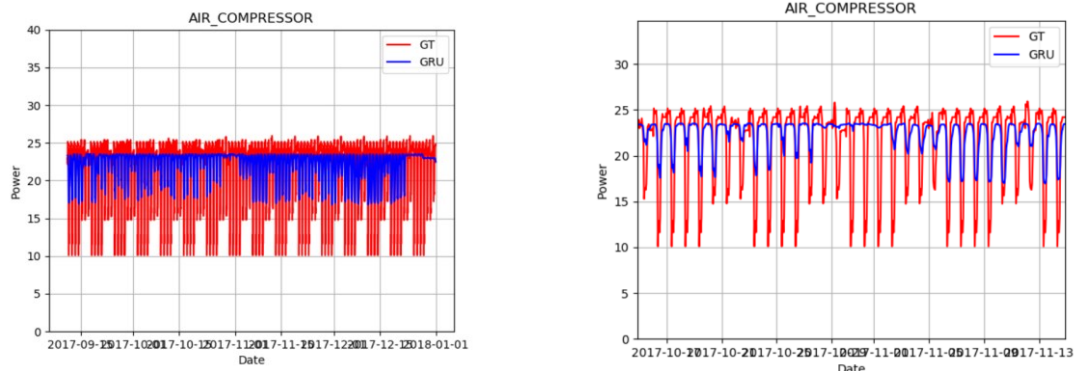


Figure 87 – Disaggregation results of the Air Compressor using GRU architecture.

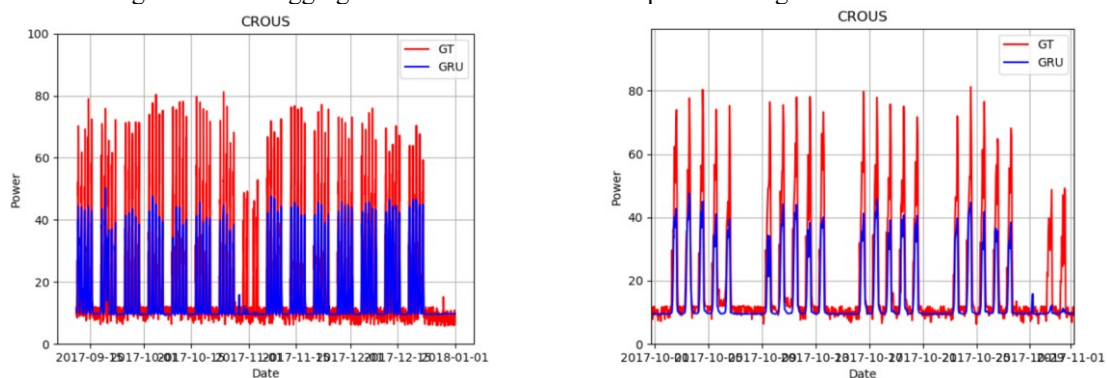


Figure 88 – Disaggregation results of the CROUS using GRU architecture.

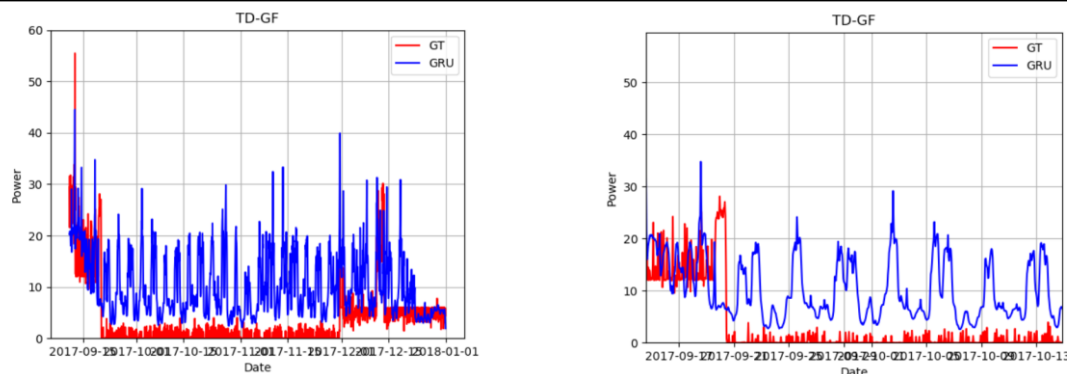


Figure 89 – Disaggregation results of the TD-GF using GRU architecture.

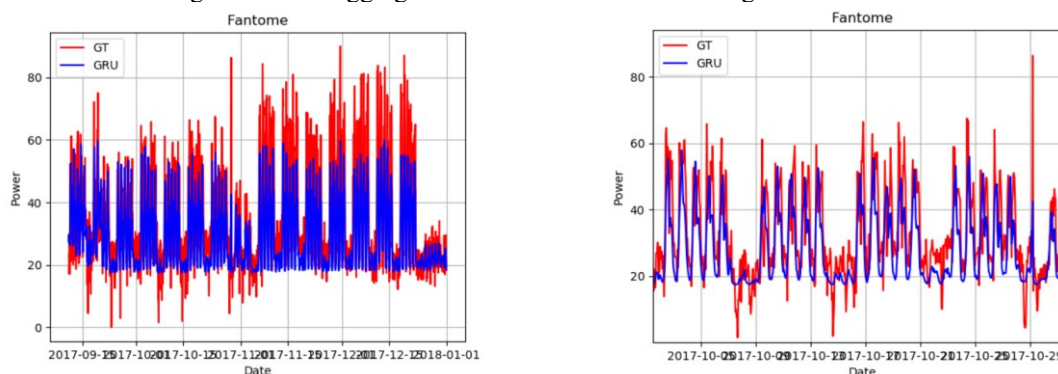


Figure 90 – Disaggregation results of the “Fantome” load using GRU architecture.

Table 28 – Window GRU network disaggregation results.

Load	Average power [kW]		MAE [kW]	Percentage error
	Ground Truth	Disaggregated		
Air Compressor	21,80	22,13	2,21	1,51%
CROUS	19,32	15,94	5,46	17,49%
TD-GF	3,44	10,17	7,75	195,64%
Fantôme	33,18	28,84	7,43	13,08%

Despite presenting higher percentage error when compared to the LSTM approach, it can be said that the Window GRU algorithm improved the results. It can be confirmed by visualization of the previous figures along the lower MAE exposed in the previous table.

### 6.2.5 Sequence-to-point

For the sequence-to-point approach, detailed in the item 5.3.2.3, a batch size of 1024 and 50 epochs were used as parameters to configure the training phase. The results obtained are illustrated in the next figures and in the following table.

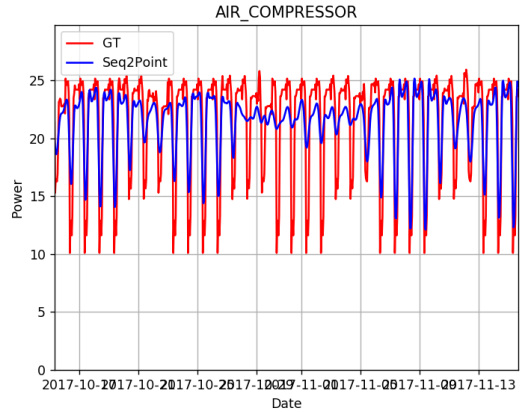
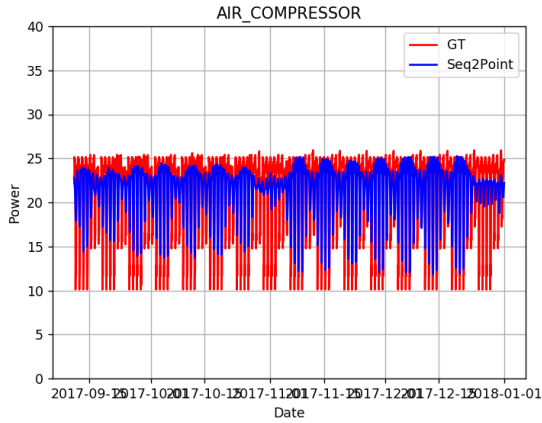


Figure 91 – Disaggregation results of the Air Compressor using Sequence-to-point.

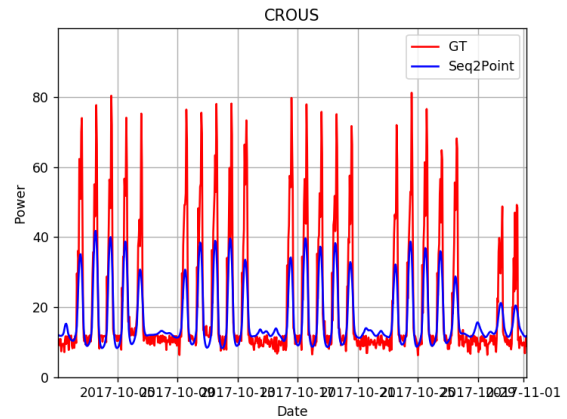
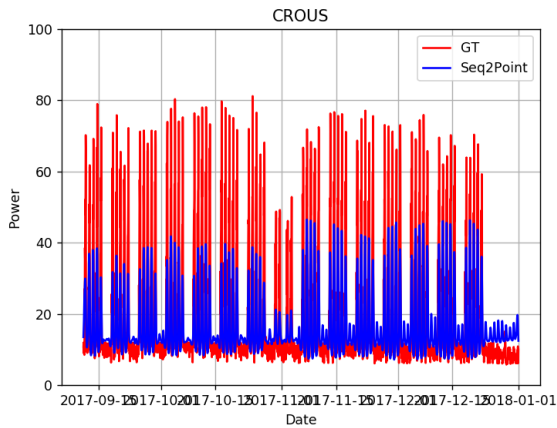


Figure 92 – Disaggregation results of the CROUS using Sequence-to-point.

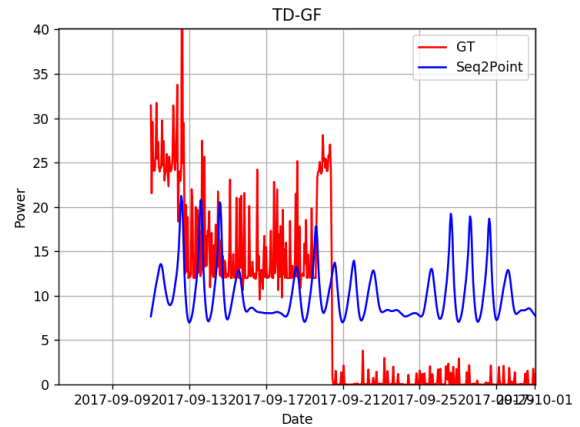
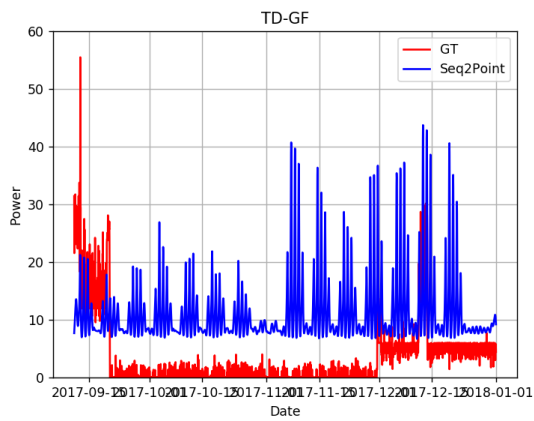


Figure 93 – Disaggregation results of the TD-GF using Sequence-to-point.

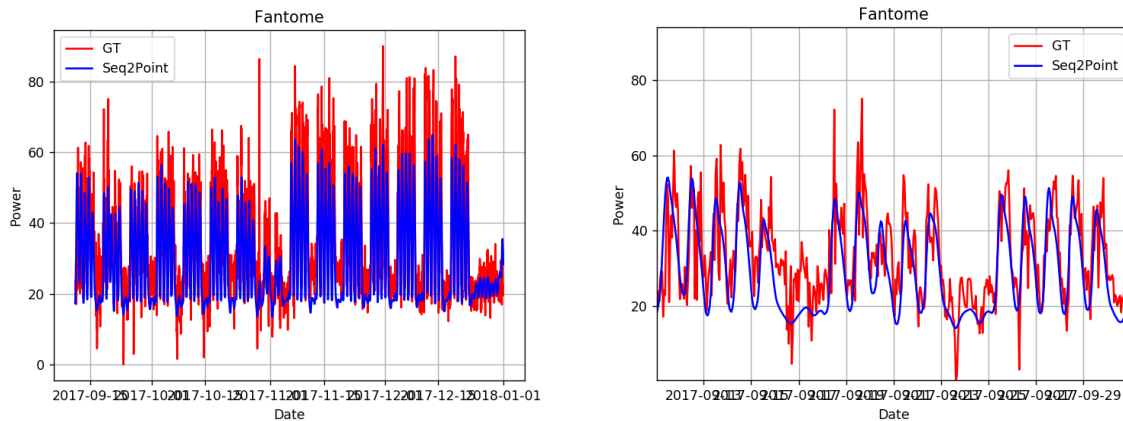


Figure 94 – Disaggregation results of the “Fantôme” load using Sequence-to-point.

Table 29 – Sequence to point disaggregation results.

Load	Average power [kW]		MAE [kW]	Percentage error
	Ground Truth	Disaggregated		
Air Compressor	21,80	21,47	1,95	1,51%
CROUS	19,32	17,09	6,17	11,54%
TD-GF	3,44	11,02	9,04	220,35%
Fantôme	33,18	29,45	7,04	11,24%

The results issued from the Sequence-to-point application to the dataset presented in this chapter were slightly improved compared to the Window GRU approach. The percentage error was lower for the CROUS and Fantome loads and the same for the Air Compressor. Considering the MAE, the Sequence-to-point presented lower values for the Air Compressor and the Fantome load, while the CROUS was slightly higher, endorsing the similar performance between this approach and the previous one.

### 6.3 Conclusions

This chapter aimed at applying the NILM methods described in the previous chapter to the GreEn-ER dataset. The results obtained show that the energy disaggregation in the tertiary sector is not trivial. The number of loads, their pattern and their share relative to the overall consumption decrease the performance of energy disaggregation algorithms. In addition, neural network estimates are computationally demanding, making the application of neural network-based algorithms virtually impossible for embedded systems or even standard computers when dealing with high number of loads.



Because of that, the choice of target loads is of utmost importance. It reduces the number of loads, disregarding less important loads, reduces the computation time and improves the performance of the techniques.

The comparison between the disaggregation methods used indicate that techniques based on artificial neural networks were better suited to the case study loads. This is explained by the nature of many loads present in buildings outside the residential environment. These are loads of higher power, with pattern not always with well-defined states. Even so, one can highlight that among the techniques based on finite states, the FHMM has an advantage when compared to the Combinatorial Optimization. Furthermore, it can be said that the FHMM can deliver satisfactory results considering simple pattern loads, such as the air compressors, in a much faster time than the ANN-based algorithms. This indicates that, besides the best ANN based algorithm, the Sequence-to-point, the FHMM could also be used in a tertiary building as disaggregation technique under certain circumstances.

## References

- [1] EPRI. Low-cost nialms technology: Market issues and product assessment. Electric Power Research Institute, Palo Alto, California, TR-108918-V1, 1996.
- [2] EPRI. Low-cost nialms technology: Market issues and product assessment. Electric Power Research Institute, Palo Alto, California, TR-108918-V2, 1996.
- [3] L. K. Norford and S. B. Leeb. Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms. *Energy and Buildings*, 24(1):51–64, 1996
- [4] V. A. Smith, L. Norford, and S. Leeb. Final report compilation for equipment scheduling and cycling. California Energy Commission.< [http://www. archenergy. com/cec-eeb/P2-Diagnostics/P2-Diagnostics Reports P, 2, 2003](http://www.archenergy.com/cec-eeb/P2-Diagnostics/P2-Diagnostics Reports P, 2, 2003).
- [5] A. Meier, D. Cautley, Practical limits to the use of non-intrusive load monitoring in commercial buildings, *Energy and Buildings*, Volume 251, 2021, 111308, ISSN 0378-7788, <https://doi.org/10.1016/j.enbuild.2021.111308>.
- [6] N. Batra, O. Parson, M. Bergés, A. Singh and A. Rogers. A comparison of non-intrusive load monitoring methods for commercial and residential buildings. *ArXiv*, abs/1408.6595. 2014
- [7] P. Bandeira de Melo Martins. *Non-intrusive industrial load monitoring on a factory in Brazil*/Pedro Bandeira de Mello Martins. – Rio de Janeiro: UFRJ/COPPE, 2020. XIV, 83 p.: il.; 29, 7cm
- [8] Z. Ling, Q. Tao, J. Zheng, P. Xiong, M. Liu, Z. Xiao, W. Gang. A Nonintrusive Load Monitoring Method for Office Buildings Based on Random Forest. *Buildings*, 2021, 11, 449. [https://doi.org/10.3390/ buildings11100449](https://doi.org/10.3390/buildings11100449)
- [9] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring. In: 5th International Conference on Future Energy Systems (ACM e-Energy), Cambridge, UK. 2014. DOI:10.1145/2602044.2602051. arXiv:1404.3878.
- [10] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring". In: NILM Workshop, Austin, US. 2014 [pdf]
- [11] J. Kelly, N. Batra, O.r Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, M. Srivastava. Demo Abstract: NILMTK v0.2: A Non-intrusive Load Monitoring Toolkit for Large Scale Data Sets. In the first ACM Workshop On Embedded Systems For Energy-Efficient Buildings, 2014. DOI:10.1145/2674061.2675024. arXiv:1409.5908.
- [12] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson. Towards reproducible state-of-the-art energy disaggregation. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19).

---

Association for Computing Machinery, New York, NY, USA, 193–202. 2019.  
DOI:10.1145/3360322.3360844

## 7 Energy Audits Using Energy Disaggregation – The GreEn-ER Compressed Air System Case

An energy audit is an important tool towards the prospection of energy savings potentials. This analysis usually uses on-site data measured by the auditors. However, the time available for these measurements is limited and may not include some modes of operation of certain appliances. One example of that is the quantification of compressed air leaks, which can be done by estimating the flow rate during a no compressed air consumption period. Nevertheless, these periods often do not coincide with auditors' schedule, issue that could be addressed by using historical data. However, historical data from energy management systems usually are only available for global consumption, and rarely for individual appliances. In this context, a NILM approach would be helpful to enhance energy audits carrying analysis of modes of operation not included in the on-site measurements.

Hence, Chapter 7 discusses the possibility of using NILM techniques to enhance energy audits by estimating the leakage in the GreEn-ER compressed air system using energy disaggregation. The results obtained show that it is possible to use energy disaggregation to estimate compressed air leaks in the context of an energy audit.

### **Chapter Contents**

<b>7.1 Disaggregation and Energy Audits Overview</b> -----	<b>148</b>
<b>7.3 Energy Audits Overview</b> -----	<b>149</b>
<b>7.3 Case Studies – Estimating Compressed Air Leakage</b> -----	<b>152</b>
<b>7.4 Conclusions</b> -----	<b>180</b>

## 7.1 Disaggregation and Energy Audits Overview

A first step towards the reduction of energy consumption is the realization of an energy audit. This audit is a detailed inventory of the energy performance of the systems in a residential, tertiary or even industrial environment. An audit makes it possible to become aware of the quality of the energy installations and the daily behaviors and must provide personalized and quantified advice to consume energy in a more rational way. The European Directive 2012/27/EU, effective in France since 2013, already mandates all sectors of companies to perform an energy audit, every four years. [1]

A successful energy audit seeks first to identify the major energy consumers, where the biggest energy savings potentials lies. In this context, compressed air systems play a key role in many tertiary buildings. These systems usually are among the major energy consumers in a facility. Because of that, even small relative potential savings in compressed air systems may represent a big reduction in the consumption. In this circumstance, this chapter aims at enhancing energy audits in a tertiary building environment by using NILM (Non-Intrusive Load Monitoring) techniques [2] to estimate the compressed air leakage.

The evaluation of this possibility is done by evaluating the estimation of compressed air leakage in an adapted dataset from the GreEn-ER building. The first section details how an energy audit usually unfolds. Afterwards, a method to estimate the flow rate from the electric power of a rotary-screw fixed speed air compressor, as one installed in GreEn-ER facilities is presented.

Subsequently three case studies are exposed. The first one uses synthetic data to evaluate the possibility of using a NILM technique, with just one week of training period, to estimate the power, and thus the flow rate of a rotary-screw fixed speed air compressor, in a period in which the equipment presents a different behavior. With the promising results obtained from this analysis real data from an air compressor installed in the GreEn-ER building was used as replacement of the synthetic data. Using the original data, sampled every 10 minutes the power associated to both load and unload states were not correct, probably because the air compressor changes its states with higher rate than 10 minutes sampling. This fact makes flow rate estimation inaccurate, and also the NILM approximation. To work around this problem the data was upsampled to 1 minute sampling interval, maintaining the same consumed energy but using the correct power associated to each state. The results in this case, was also promising.

## 7.2 Energy Audits Overview

Energy audits are one way to obtain accurate and objective assessments of how to achieve savings. An energy audit is a process by which a building is inspected and analyzed by an experienced technician to determine how energy is used in it, with the goal of identifying opportunities for reducing the amount needed to operate the building while maintaining comfort levels [3].

There are several types of energy audits, and the level of complexity and detail of the analysis performed classify them. The first and less complex type of energy audit is the Benchmarking Audit. It performs a detailed preliminary analysis of the energy consumption and its cost, relying on utility bills, determining benchmark indices, like the ratio between the energy consumption and the surface area in a determined period of time, usually a year [3].

The second type is the walk-through audit. It consists in a quick tour in the facility to visually inspect the target systems. This may include the analysis of energy consumption patterns and provide comparisons to average benchmarks for similar facilities. When the inspection of the target systems show promising savings potentials, this audit can lead to a more complex audit later [3].

The standard audit, seeks to quantify energy consumption and losses by performing a detailed analysis of the performances of several energy systems. This analysis usually includes on-site measurements, historical data collection and testing to determine the efficiency in the energy analyzed systems. Specific energy engineering calculations are applied to determine efficiencies and calculate energy and financial savings based on improvements and changes to each system. As the name suggests and because of its cost-benefit, it is the most common type of energy audit performed. However, when historical data is not available, and the audit relies on on-site measurements, a photo of the operating conditions and the extrapolation for other operating points of the systems may be laborious to obtain [3].

To address this problem, there is a more complex type of energy audit. It relies on computer simulations to predict the consumption that was not contemplated in the on-site measurements phase. The goal is to build a base for more consistent comparison with the actual energy consumption of the analyzed facility. This baseline is then used to compare with the performances achieved with improvements and changes tested in the simulation

environment. Because of the time involved in collecting data and setting up an accurate simulation model, this is the most expensive level of energy audit [3].

As the Standard Energy Audit is the most common type, the work developed in this thesis have been based on it. This procedure can be divided in three phases. The first phase concerns the collection of all necessary data for the efficiency analysis of the evaluated systems. Visual inspections, technical data from catalogs, historical data from the supervisory system, when available, and field measurements are examples of the collected data. This fieldwork consists of taking measurements of various quantities (electrical, thermal, luminous, physical etc.) necessary to determine the efficiency of the evaluated systems. These measurements should be made following standard technical procedures for each type of system evaluated, using reliable meters adapted to each type of installation and system.

One way to perform this data collection is by incorporating the team into the environment to be evaluated. In this manner the team can experience the daily operation of the facility. However, besides the fact that the auditors' time to perform this task is limited, usually a few days or weeks, the ideal is that their activities have as little impact as possible on the normal operation of the systems being evaluated. Because of this, some operating modes of certain equipment may not be measured during the auditors' data collection period.

The second phase consists in analyzing the collected data and determining the consumption and efficiency of the evaluated systems. To perform this task, the data collected in the preceding phase is applied in procedures specific to each system. In this stage, possible opportunities to promote energy sobriety can also be identified.

The last phase consists in proposing improvements for the reduction of energy consumption, or even the replacement of specific equipment by a more efficient or cheaper one. Replacement by more efficient equipment, changes in operating procedures or the installation of new components that promote more rational energy are among the most common solutions. All this information is then made available to the customer in the form of a report so that he can appreciate the alternatives presented and choose whether or not to make these improvements. Figure 95 illustrates the flowchart process of a standard energy audit.

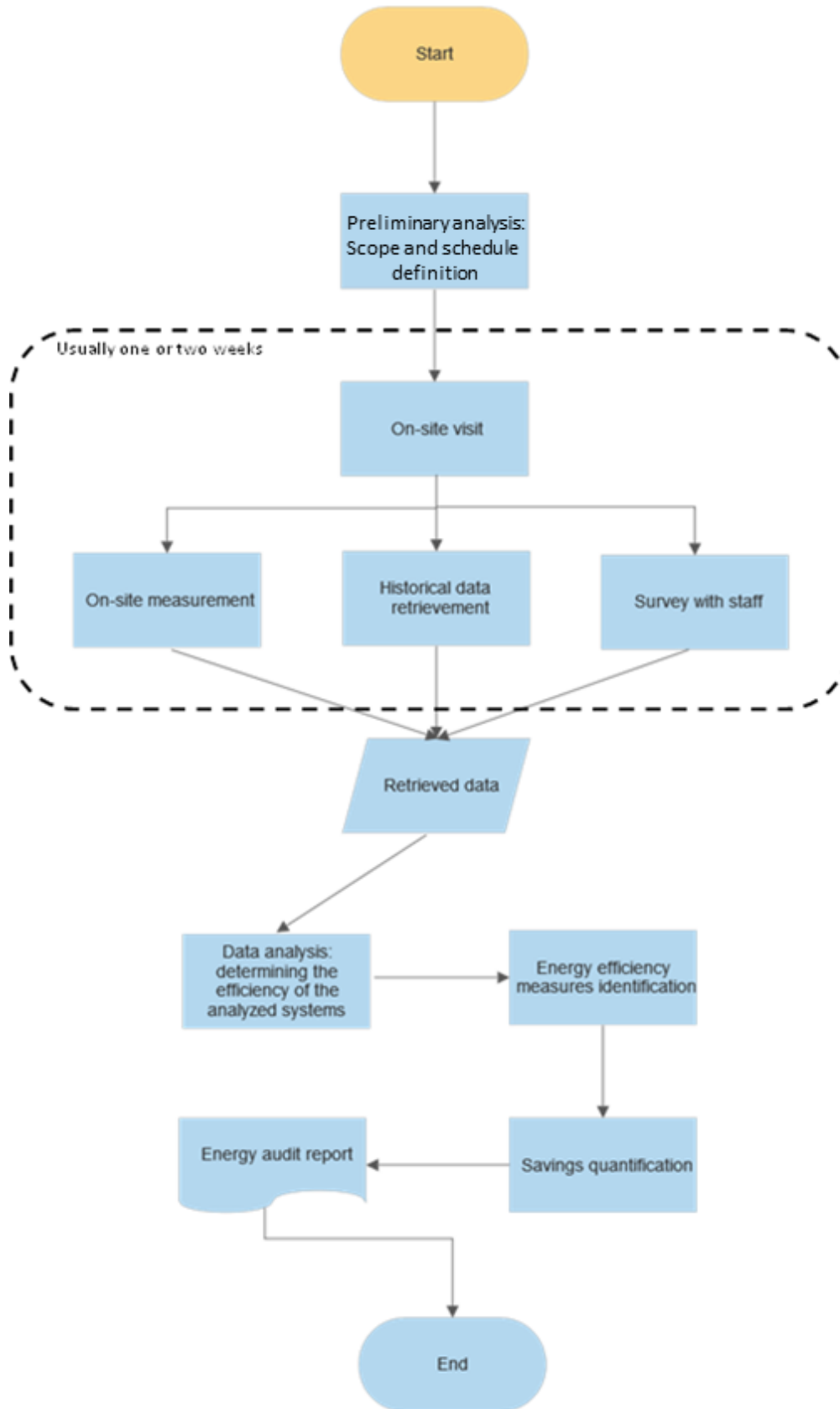


Figure 95 – Flowchart process of a standard energy audit.



Usually, large energy consuming systems are the main targets of an energy audit, because even a small absolute reduction in the consumption of these systems can represent large absolute savings. In buildings, lighting systems, HVAC (Heating Ventilation and Air Conditioning) and the envelope are usually emphasized. In industrial environments, besides those mentioned for buildings, analyses in water pumping systems, motors, boilers and compressed air systems are common.

### **7.3 Case Studies – Estimating Compressed Air Leakage**

Compressed air systems usually are among the major energy consumers in a facility in which these systems are present, whether in a building or in an industrial environment. The use of a variable speed air compressor instead of a fixed-speed one, and the quantification and repair of leaks are among the most common solutions to improve the efficiency and achieve savings in a compressed air system.

Leaks are a significant cause of wasted energy in a compressed-air system and can develop in many parts of a compressed air system. The most common problem areas are couplings, like hoses, tubes, fittings pipe joints, quick disconnects, filters, regulators, condensate traps, valves, flanges and other point-of-use devices [4]. Because of the similar physical characteristics with other gases, the compressed air leakage quantification and detection can be dealt in an analogous way as other gases. Several methods have been developed to detect and quantify fluid leaks in pipelines. They can be classified into biological, hardware and software techniques [5].

The biological ones rely on empirical methods using sensorial perceptions, such as the hearing, smell, sight of specially trained staff. The hardware techniques use numerous equipment to enhance the sensorial perception of the staff. In the case of compressed air systems, the most used are ultrasound devices [6] [7] and infrared cameras [8]. Although it is considered as the industry standard and best practice, the ultrasonic leak detection is limited to the application in short distances [9] and requires highly trained operators. Moreover, due to the compressed air expansion process in the location of the leak, a temperature gradient is created, making it possible to observe it from infrared images. However, these techniques also require the measurement of the openings through which the compressed air escapes from the pipeline to quantify the leaks and thus the potential savings from their repair. The measurement of those openings may be impractical for the auditors. Software-based methods

use flow, pressure, temperature and other data to estimate the leaks. These methods are usually based on the analysis, performed by an automatic algorithm, of both pressure and flow rate data, when available [5] [10].

An alternative to quantify the compressed air leakage is to estimate the air compressors flow rate during a no compressed air consumption period, such as a weekend or a holiday [11]. During these periods, all the compressed air end use equipment should be turned off. In this scenario all the compressed air is directed to feed only the leaks. This estimation may be done by measuring the input power of the air compressor and correlating it to its flow rate. However, this technique implies that the auditors must be on site during a no compressed air consumption period for the monitoring of the air compressors. In an industrial environment, for instance, a no compressed air consumption period is rare, and it may not coincide with the energy audit planning. An alternative to go around this problem would be collecting logged data from the compressed air system. However, it is unlikely that the input power of the air compressors, or even the flow rate, is monitored individually. Although it is improbable to have the air compressors monitored individually, logged data from the global consumption is often available.

As the correlation between the input power and the flow rate of an air compressor is an intrinsic characteristic of the equipment, it remains the same regardless the system's operation mode. Thus, if one could extract the air compressors input power from the global consumption, the leakage estimation could be performed even without the presence of the auditors on site during a no compressed air consumption period. This would enhance the quality of the analysis performed during the energy audit with a smaller cost when the other options for this analysis are impractical.

The load curve extraction of an equipment from the global consumption can be done using NILM (Non-Intrusive Load Monitoring) techniques [2] under certain circumstances. Thus, this work aims at investigating the possibility of using NILM methods to estimate the compressed air leakage in a tertiary building environment and to calculate the potential savings with the leaks repair in the context of an energy audit.

NILM techniques were already used to detect leaks in compressed air systems. In his master thesis [12], Piber used a stationary device to measure the global consumption of a warship. With embedded NILM algorithms, this device was also capable to extract the consumption of some loads, such as vacuum-assisted sewage collection system, a low-

pressure compressed air system. Analyzing the operating schedule, he was able to detect the presence of leaks in the compressed air system. However, he did not mention the sampling rate neither the training period used to train the NILM algorithms. In addition, as a stationary device was installed, a permanent change in the facility were done. Therefore, it is important to highlight that the proposed technique is non-intrusive and does not imply any permanent changes in the system, such as installing flow meters in the pipeline, or stationary power meters. In addition, the NILM technique to enhance energy audits was already discussed by Berges et al. [13], but it was limited to the application in a residential environment.

Summarizing, the idea is to use input power measurements that could be retrieved using portable power analyzers, for instance, to train the algorithm and logged data from the global consumption to extract the air compressors power consumption. This estimation, if performed in a period with no compressed air consumption (weekends, vacations, holidays...), represents the leakage present in the grid. If the amount of the leakage is known, the estimation of the savings achieved with the repair of some, or the totality, of the leaks is feasible. Thus, with input power data, during both normal operation and no compressed air consumption periods, and the data sheets of the equipment it is possible to estimate the potential saving by repairing the leaks present in the compressed air system. The procedure to estimate the potential savings with the repair of the compressed air leakage is illustrated in Figure 96.

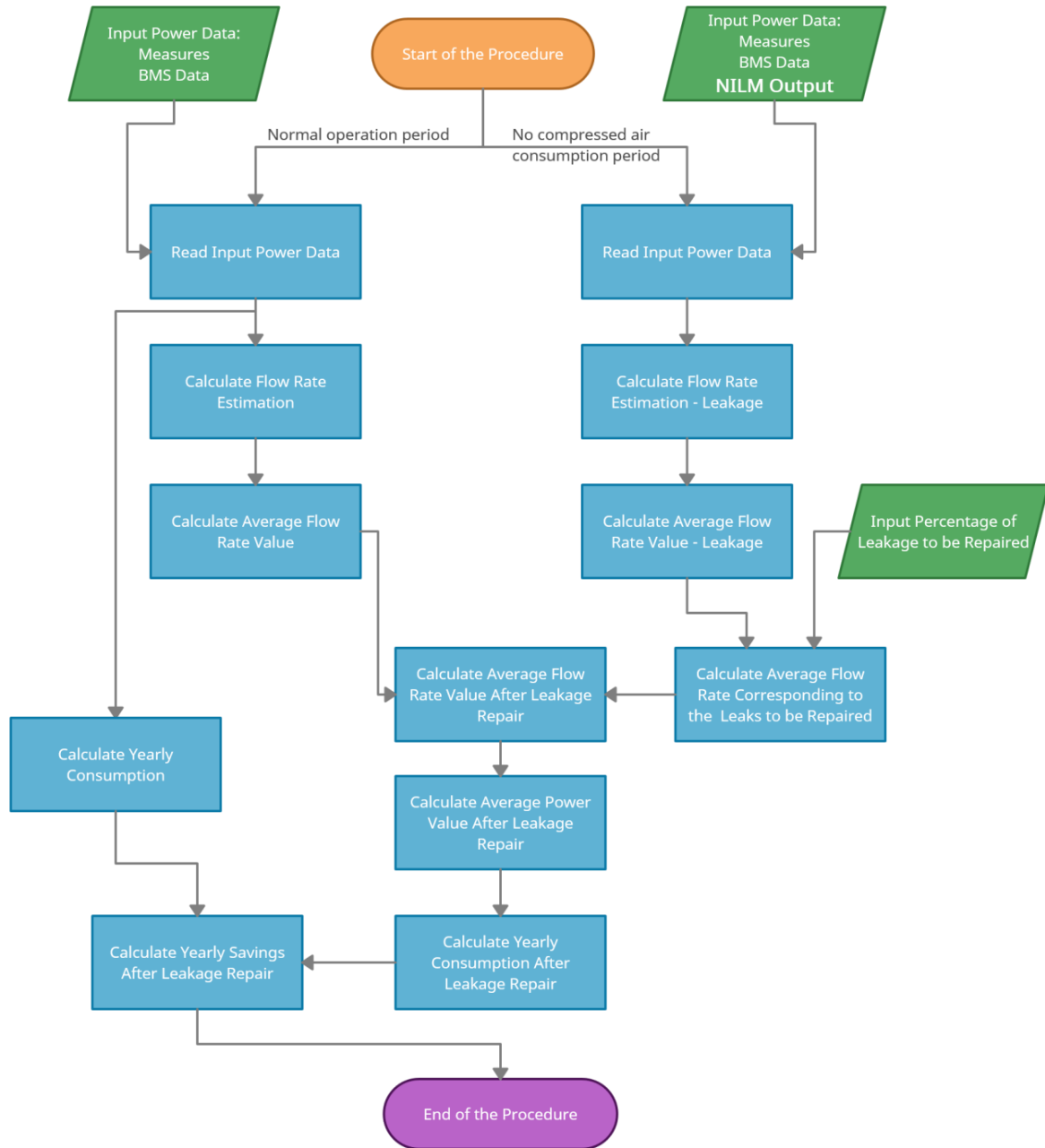


Figure 96 – Potential savings estimation with the repair of compressed air leakage.

### 7.3.1 Estimating Compressed Air Flow From Input Power Measurements

Tables (fixed-speed) and curves (variable-speed) provided by the air compressors manufacturers make the correlation between the input power and the compressed air flow rate possible. When this information is not available, it is possible to measure the flow rate by measuring the air velocity at various points on the intake pipe cross-section and integrate these measurements in the pipe cross-section area, with an anemometer, for example. For the variable-speed compressor, numerous measurements should be done in several operating

points, in order to determine a correlation curve between power and flow rate. The flow rate estimation by correlating the input power of the air compressors and their flow, enables also the determination of load curve of compressed air in a facility, in the absence of a flow meter.

Fixed-speed air compressors operate at full capacity, delivering rated flow, until the pressure set point is reached. At this point, the compressor unloads, operating at minimum power, to maintain internal pressure, while producing no air to the network. Because of this behavior, this type of compressor is also called modulating compressor. Usually, it varies between two states, the load and the no load one. These two states typically have very distinguishable input power associated, in a way that it is not difficult to identify them.

Therefore, to estimate the flow of a fixed-speed air compressor it is enough to well identify its load state and correlate it with the compressor's flow rate capacity, measured or obtained by manufacturer tables. For the no load state, zero flow is assigned. An example of a performance table of a rotary-screw fixed speed air compressor operating at rated pressure provided by the manufacturer is presented in Table 30, while an example of the association between the power and the flow rate is illustrated in Figure 97.

Table 30 – Manufacturer performance table of a Fixed-speed air compressor [14].

Model	GS30B10
Manufacturer	BelAir
Rated Capacity at Full Load Operating Pressure [m <sup>3</sup> /h]	294
Full Load Operating Pressure [bar]	8
Drive Motor Nominal Rating [hp]	40
Total Package Input Power at Zero Flow [kW]	6
Total Package Input Power at Rated Capacity and Full Load Operating Pressure [kW]	33.8

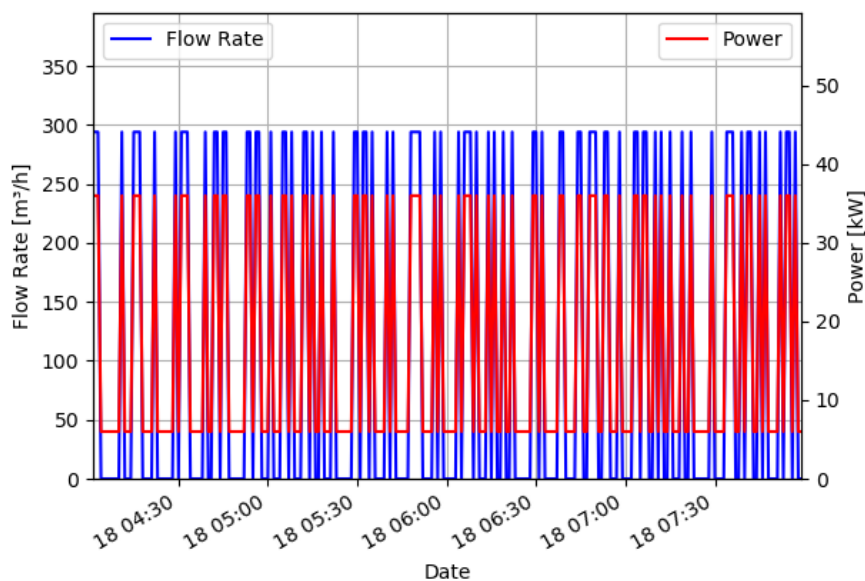


Figure 97 – Example of association between power and flow rate of a rotary-screw fixed speed air compressor.

Another way to estimate the flow rate from the power measurements is by the use of the empiric Equation 26, in which  $P_{avg}$  represents the average power,  $P_{load}$  and  $P_{unload}$  are the power associated to both load and unload states, respectively.  $Q_{avg}$  stands for the average flow rate, while  $Q_{rated}$  is rated flow rate of the air compressor at the load state. This equation allows to calculate the average flow rate of a rotary-screw fixed speed air compressor.

$$Q_{avg} = \frac{(P_{avg} - P_{unload})Q_{rated}}{(P_{load} - P_{unload})} \quad \text{Equation 26}$$

### 7.3.2 Estimating Compressed Air Leakage Using NILM Techniques

In order to perform an estimation of compressed air leakage in a facility using NILM techniques, consumption data from a rotary-screw fixed-speed air compressor were extracted from the GreEn-ER dataset [16]. The computational code was developed in Jupyter Notebooks, an environment in python that combines code, text and images. It is also available in open source [17]. The original data, especially the global consumption, contain some data quality problems, such as the presence of outliers. These anomalous values were identified using the forecast error method, presented in Chapter 4 [18] and corrected.

Two approaches were evaluated. In the first one, a synthetic dataset of a fixed-speed air compressor was created based on a real one. It was inserted into the GreEn-ER dataset, in substitution of the real air compressor that exists in the building. Afterwards, real data of the air compressor from the GreEn-ER building were used, extracted from the building's management system.

#### 7.3.2.1 *Applying NILM Algorithms to Quantify Compressed air Leaks Using Synthetic Data*

In the context of an energy audit, it is unusual that the auditors are on site to measure the air compressor consumption during no compressed air consumption periods. Additionally, measuring and logging the power of individual appliances is not the standard in most facilities, either industrial or even tertiary buildings. However, it is more common for global consumption to be measured and logged by an energy management. Therefore, one alternative is to use NILM algorithms to estimate the power of such appliances from the global consumption, during these periods that the auditors are not on site.

It is important to remember that the functioning of a fixed-speed screw air compressor, the most common used in most facilities around the world, is slightly different during normal operation periods and no compressed air consumption periods. These air compressors usually operate in two well defined states, called load and unload. During the load state, the compressor operates at its rated capacity delivering maximum air flow rate, at rated power. On the other hand, during the unload state, the air compressor delivers zero air flow rate, keeping a residual consumption. To adjust the air production to the air consumption, the air compressor modulates between these two states, controlled by the grid's air pressure. Hence, during normal operation periods, a well-sized air compressor tends to operate most of the time at the load state, while during periods of reduced, or zero, compressed air consumption, the unload state would be more present, creating two different operation modes.

As stated before, to estimate the compressed air leakage, it is sufficient to well estimate the operation the air compressor during a no compressed air consumption period. However, during the on-site measuring campaign by the auditors, only the normal operation mode would be available for data collection and training an eventual NILM algorithm. Thus, in order to assess the possibility of using NILM techniques to estimate the air compressor functioning during a no compressed air consumption period having only data from the normal operation mode for training, a well-behaved synthetic dataset was created. It was based on the specifications of a real air compressor, the GA18+-100 from Atlas Copco, which specifications are shown in Table 31. This dataset was separated into two different periods, one regarding normal operation and the other regarding a period with be no compressed air consumption. The data created were then inserted into the GreEn-ER dataset in substitution to the actual air compressor present in the building.

It was created one year worth of synthetic air compressor data, with one hour sampling interval. Firstly, a normal period was created. The time step chosen was divided in ten equal steps and the duration of a load period was taken randomly between five and ten steps. At the end of the series, simulating a period of no-production of one week, the duration of a load period was also taken randomly, but this time with duration between zero and five steps. White noise was also inserted in the data. The following figure shows the generated data, and a zoom in detailing a normal operation period and a no-production one. At the same time Table 32 presents the average power in each period.

Table 31 – Manufacturer performance table of a Fixed-speed air compressor. [19]

Model	GA18+-100
Manufacturer	Atlas Copco
Rated Capacity at Full Load Operating Pressure [m <sup>3</sup> /h]	224.3
Full Load Operating Pressure [bar]	7
Drive Motor Nominal Rating [hp]	25
Total Package Input Power at Zero Flow [kW]	5.4
Total Package Input Power at Rated Capacity and Full Load Operating Pressure [kW]	23.5

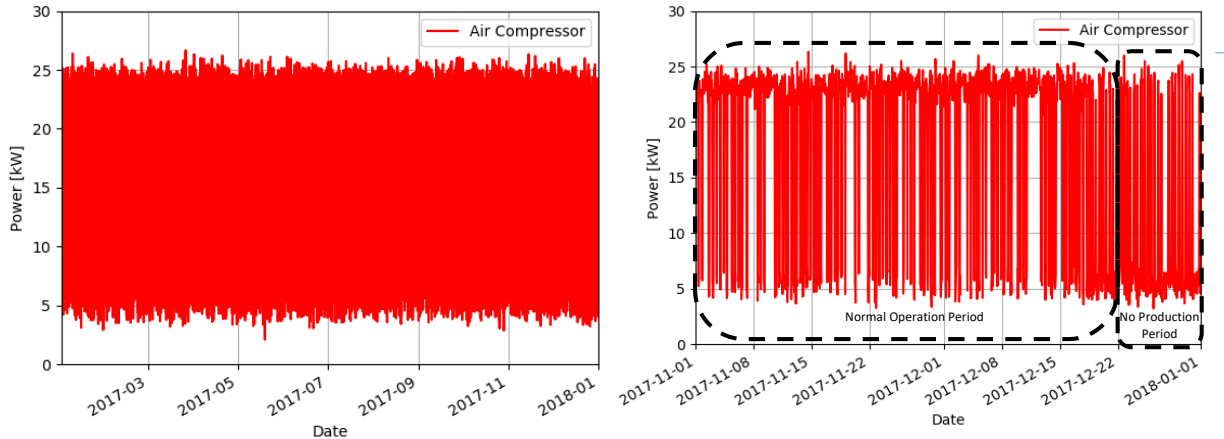


Figure 98 – Power data for the Air Compressor.

Table 32 – Average air compressor power and flow rate during the different operation periods

Evaluated Period	Average Power [kW]
Complete	18.41
Normal Operation	18.63
No-production	9.71

### 7.3.2.1.1 Leakage Estimation from Measurements – Synthetic Data Study Case

In order to assess the performance of NILM techniques in the task of quantifying compressed air leakage, it is first necessary to establish the ground truth. Hence, using the procedure presented in the section 7.3.1 the flow rate load curve was then estimated based on the air compressed power for the whole available data. Figure 99 presents a seven days rolling average of the estimative of the compressed air flow rate and the air compressor power. It is possible to visualize normal operation and no-production period, at the end of the series. Table 33 exposes the average power and estimate flow rate and the standard deviation for three periods: the complete dataset, the normal operation period and the no-production period.



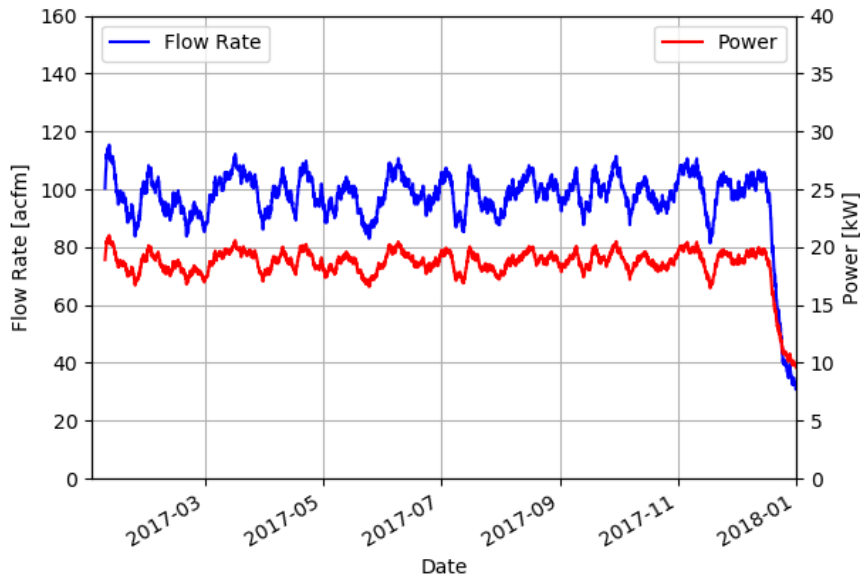


Figure 99 – Seven days rolling average of air compressor power and flow rate.

Table 33 – Average air compressor power and flow rate during the different operation periods

Evaluated Period	Average Power [kW]	Average Flow Rate [m <sup>3</sup> /h]
Complete	18.41	162.99
Normal Operation	18.63	165.77
No-production	9.71	54.03

According to the data presented in the table above, the average flow rate during the no-consumption period, which denotes the total leakage in the grid represents circa 32.6% of the average flow rate during a normal operation period. This is a typical value, compatible with compressed air leaks found in several environments. The flow rate estimations obtained by the application of Equation 26 resulted in differences smaller than 1% when compared to the results shown in Table 33.

#### 7.3.2.1.2 NILM Estimation of Compressed Air Leakage – Synthetic Data Study Case

The leakage estimation presented in the previous section was done thanks to the availability of power consumption data in a no-production period. However, that is not the typical case in the context of an energy audit. Usually, the auditors have only a few days or weeks of data that was measured by themselves. Therefore, the use of NILM techniques to estimate the power consumption of an air compressor during a no-production period, and thus the compressed air leakage, from the global consumption shows potential in this framework.

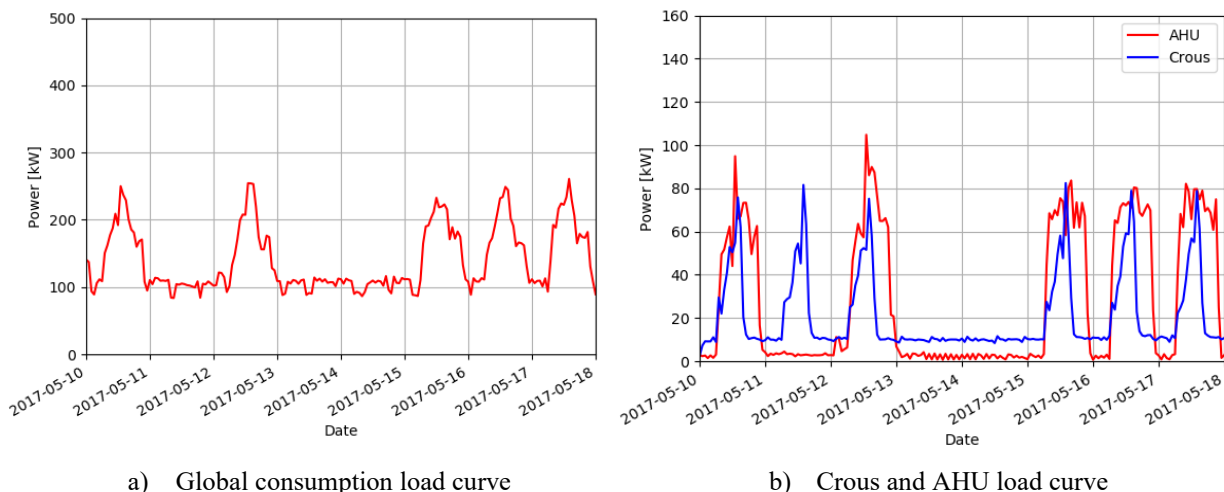
To address this task, the Factorial Hidden Markov Model (FHMM) NILM algorithm, presented in Chapter 4, was applied with the help of the NILMTK. Hidden Markov models

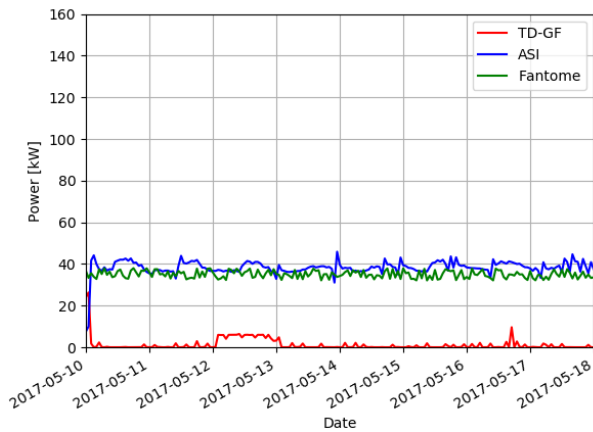
can be used in the NILM context as both supervised (using labeled data to train the algorithm) and unsupervised learning approaches, based on the requirements [15]. In the present work, the supervised approach is employed. Thus, in this approach FHMM takes some period to train the model, in order to identify power values associated to each state of operation of the target appliances.

Four major consumers of the GreEn-ER building were selected in order to create a sub dataset regarding these loads, the global consumption and the created air compressor. This sub dataset was used in the NILM task. Another load, called “fantôme” was created, corresponding to the difference between the global consumption and the sum of the other five loads. The loads selected are named as follows:

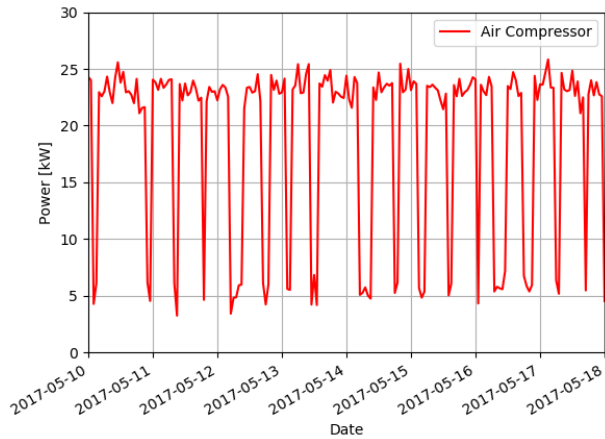
- Crous – The Crous, acronym for “Centre Régional des Œuvres Universitaires et Scolaires” load belongs to the University Restaurant, where some typical loads are connected, as dishwashers and dryers, furnaces, heaters etc.;
- ASI – The ASI load represents loads connected to the datacenter, as computers, uninterruptible power supply etc.;
- TD-GF – The TD-GF represents chillers that are responsible for the acclimatization during the summer;
- AHU – The load called AHU represents a set of 16 Air Handling Units present in the building;
- Fantome – This load is inserted as a noise, it corresponds to the difference between the main load and the sum of the other loads.

In order to simulate the context of an energy audit one typical week was selected as training period. Figure 100 illustrates the data, while Table 34 exposes its average value.





c) TD-GF, ASI and Fantome load curve



d) Air Compressor load curve

Figure 100 – Power data during the training period.

Table 34 – Average power of the loads during the training period.

Load	Average Power [kW]
Main	140.34
Air Compressor	19.06
Crous	20.94
AHU	28.33
TD-GF	1.28
ASI	38.26
Fantome	35.01

The idea behind the usage of NILM techniques to estimate the compressed air leakage is to retrieve the air compressor power from the global consumption during a no compressed air consumption period. Regarding the dataset presented in this work, the global consumption load curve, during this period, is shown in the Figure 101.

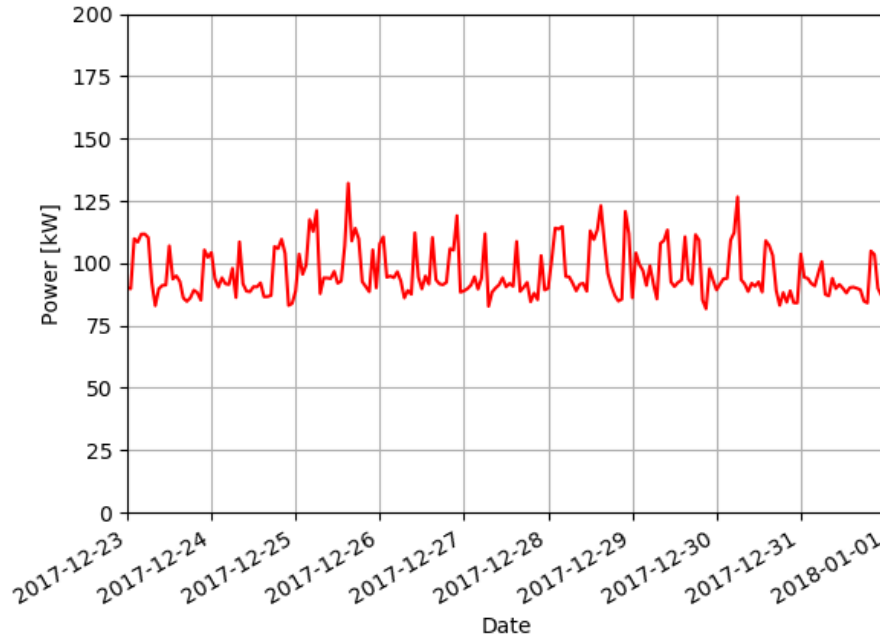


Figure 101 – Global consumption load curve in a no-production period.

With the help of the NILMTK the FHMM algorithm is then applied, which results are illustrated in the Figure 102. In this figure, GT means Ground Truth, or the consumption of the created air compressor, and FHMM stands for the results from the FHMM NILM algorithm.

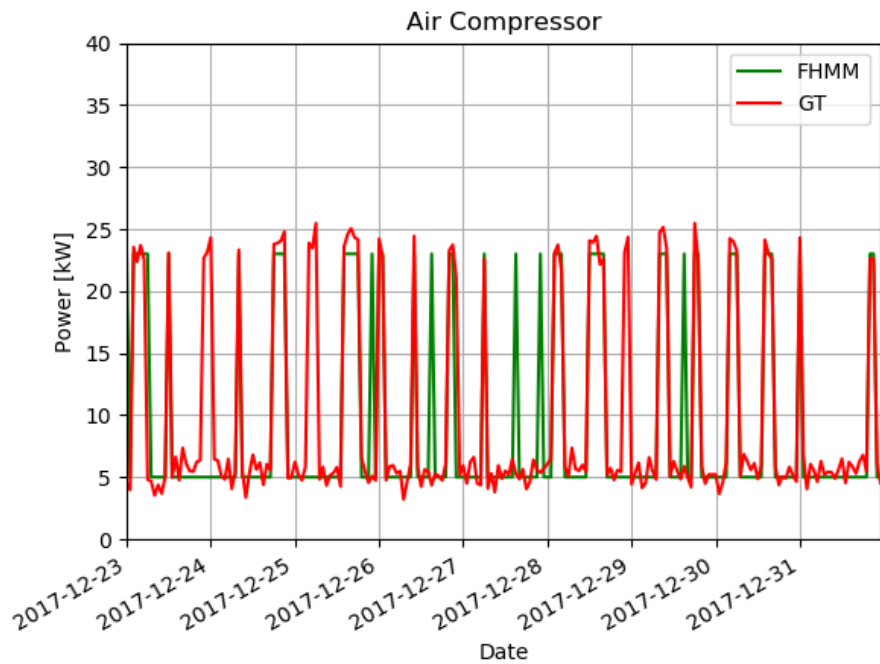


Figure 102 – Power data during the training period.

It can be seen in the figure above that the FHMM algorithm was able to retrieve most of the air compressor consumption in this example. In order to quantify the estimation of the compressed air flow rate in this period, the association between flow rate and power showed in the section 7.3.1 was applied. Then, the average compressed air flow was compared to the flow estimated to the actual compressed air power. These results are exposed in Table 35.

Table 35 – Average air compressor power and flow rate during a no-production period.

	<b>Average Power [kW]</b>	<b>Average Flow Rate [m<sup>3</sup>/h]</b>
Ground Truth	9.71	54.03
FHMM	9.16	52.18

The results presented in Table 35 suggests that, it is possible to estimate the compressed air leakage using NILM techniques in the context of energy audits, considering the dataset investigated. With only one typical week used as training the algorithm was able to estimate the compressed air flow with an error lower than 4%.

Thus, the estimative presented in the last table could be used as a result of an energy audit in the compressed air system. These results allow the auditors to estimate the energy and financial savings with the repair of the leaks and, with the estimative of the costs of the repairs, calculate the payback time. Then, the client would have the necessary information to evaluate if the investment is worth or not.

#### 7.3.2.1.3 Energy Savings With Leaks Repair – Synthetic Data Study Case

The leakage estimation performed in the previous sections allows calculating potential energy savings with its repair. As an example, this section estimates the potential savings if all of these leaks were repaired considering both estimations, the one presented in section 0 using the data considered the Ground Truth, and the other presented in the previous section performed using the results of the energy disaggregation. With the repair of all leaks, the new average flow rate can be predicted as the subtraction of the average leaks from the average flow rate of the normal operation period. Table 36 presents the prediction of the average flow rate after the repair of the leaks, considering the ground truth and the estimation from the NILM technique.

Table 36 – Average air compressor flow rate after the repair of the leaks.

Evaluated Period	Flow Rate [m <sup>3</sup> /h]	
	Ground Truth	FHMM
Normal Operation	165.77	
Leaks	54.03	52.18
<b>After repair</b>	<b>111.74</b>	<b>113.59</b>

A way to calculate average power from the average flow is by writing the Equation 26 in terms of the  $P_{avg}$ , showed in Equation 27. By obtaining the average power, it is possible to extrapolate to yearly energy consumption, and finally, the savings per year with the repair of the leaks. The average power predictions are presented in Table 37.

$$P_{avg} = \frac{Q_{avg}}{Q_{rated}} (P_{load} - P_{unload}) + P_{unload} \quad \text{Equation 27}$$

Table 37 – Average air compressor power after the repair of the leaks.

Evaluated Period	Average Power [kW]	
	Ground Truth	FHMM
Normal Operation	18.63	18.63
Leaks	9.71	9.16
<b>After repair</b>	<b>14.32</b>	<b>14.47</b>

Considering the operation schedule and the average power exposed in Table 37, it is possible to estimate the annual consumption and thus, the savings with the repair of the leaks. For the annual consumption estimation, it was considered that the no compressed air consumption period lasts one week, while the rest of the year the building remains in normal operation. During the no compressed air consumption period the air compressor would operate at the power associated to the unload state (5.4 kW), while during the rest of the year, the average power can be retrieved from Table 37. This information allows the extrapolation of the average power to annual electricity consumption and, thus, to the potential savings in the compressed air system. Table 38 exposes the results.

Table 38 – Annual electricity savings with the repair of the leaks.

Average Power	Normal operation No compressed air	Before the leaks repair	After the leaks repair	
		Ground Truth	Ground Truth	FHMM
		18.63	14.32	14.47
		9.71	5.4	5.4

<b>consumption</b>			
<b>Annual Consumption [kWh]</b>	161271.6	123956,0	125226,4
<b>Annual Savings [kWh]</b>	-	<b>37315.6</b>	<b>36045.2</b>
<b>Annual Savings [%]</b>	-	<b>23.1</b>	<b>22.4</b>

The savings predicted from the leakage estimation using the measurements represents around 23.1% of the annual consumption, while the one estimated using NILM techniques represents 22.4%, which is less than 1% of error. These results indicate that it is possible to use of NILM techniques to estimate compressed air leaks in the context of an energy audit and may enhance the savings estimations when some data are not directly available.

### **7.3.2.2 Applying NILM Algorithms to Quantify Compressed air Leaks Using GreEn-ER Data with 10 minutes sampling interval**

The results presented in the previous section, using synthetic created data of an air compressor, indicates that is possible to estimate the compressed air leakage in a grid with a rotary-screw air compressor using NILM techniques even with short period of training, like one week for example. With this background, the same procedure from the section 7.3.1 was employed, this time, using real air compressor data, issued from the GreEn-ER’s BMS. In order to simulate the context of an energy audit, two weeks’ worth of data were used in this work. The first one corresponds to a period of normal operation and will be used as training period. The other week, during a period with no use of compressed air, is used to estimate the leaks, called also test period. The data are sampled with 10-minutes sampling interval. The following sections present the estimation of the compressed air leaks using measurements data and NILM technique FHMM.

#### **7.3.2.2.1 Leakage Estimation from Measurements – GreEn-ER Data With 10 Minutes Sampling Interval Study Case**

The scheme of the compressed air production in the building is illustrated in Figure 103. In this figure, it is possible to observe, that there are two air compressors, of which one normally operates while the other is kept as system backup. Its specifications data were presented in Table 30. The power associated to the load state was identified as 18kW, while the unload state was associated with 12kW, as it can be seen in Figure 104. Additionally, Figure 105 presents the power of the air compressor in both training and test periods.

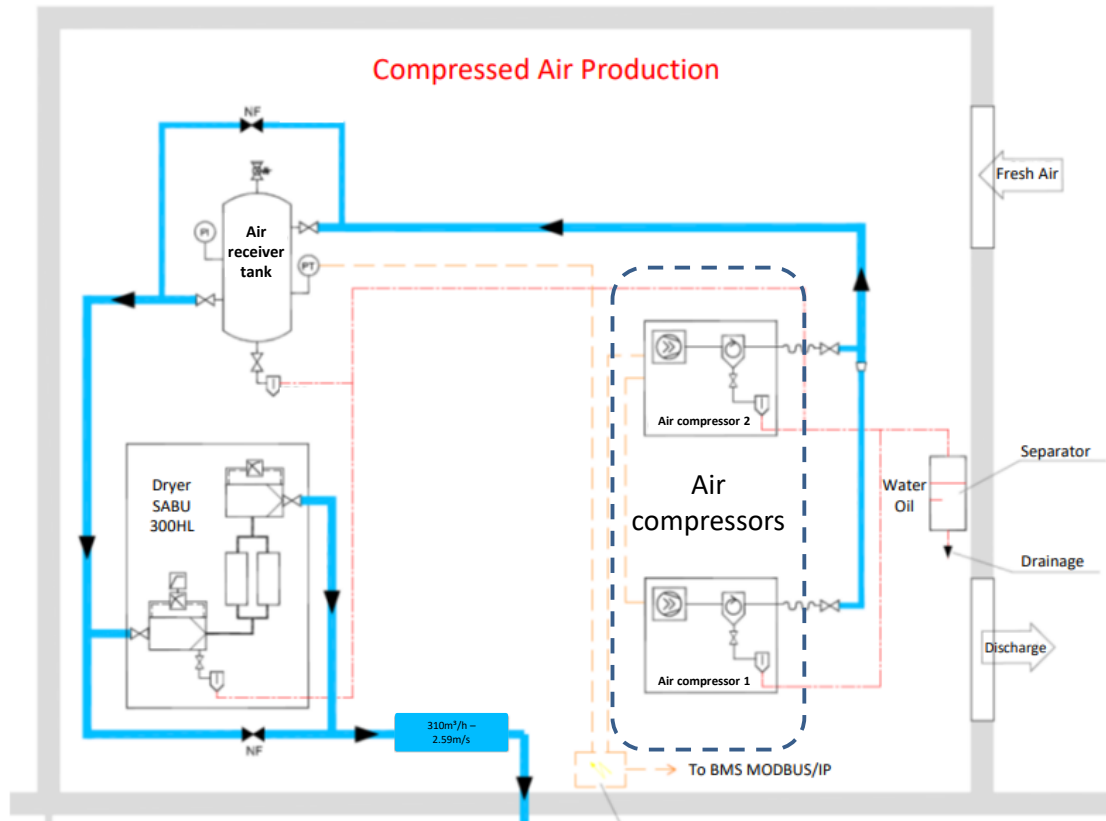


Figure 103 – Compressed air production scheme.

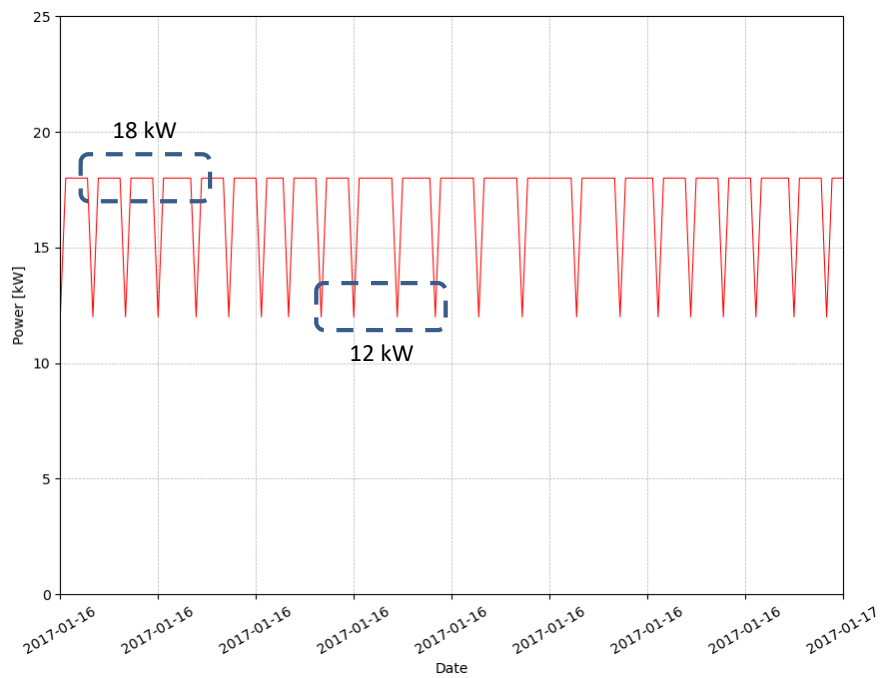


Figure 104 – Power data for the air compressor identifying the power associated to both load and unload states.



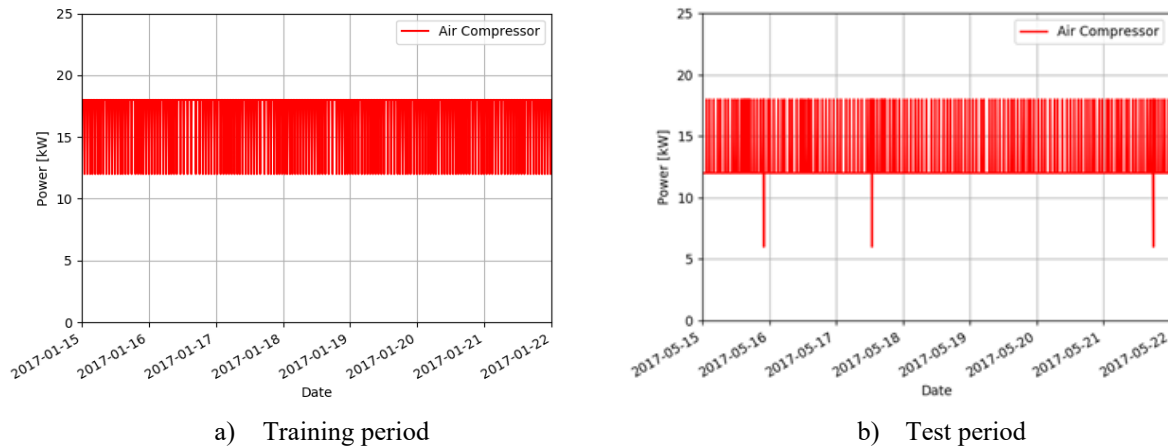


Figure 105 – Power data for the Air Compressor.

The number of samples in which each of the operating states of the air compressor were identified are presented in Table 39. In addition, using the procedure presented in section 7.3.1 the flow rate load curve was then estimated. The same table exposes the average power and estimated flow rate for the normal operation and no-production periods.

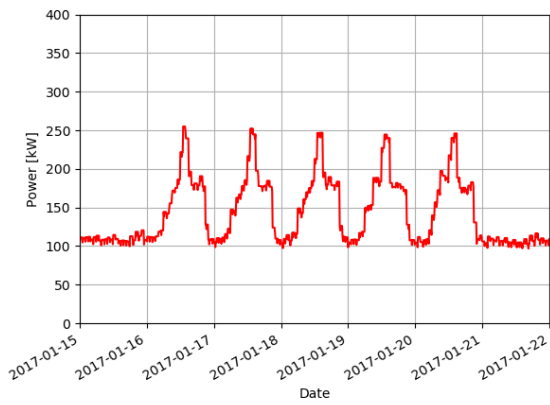
Table 39 – Average air compressor power and flow rate during the different operation periods

Evaluated Period	Number of samples		Average value	
	Load	Unload	Power [kW]	Flow Rate [m <sup>3</sup> /h]
Training Period	846	163	17.03	246.70
Test Period	140	869	12.83	40.79

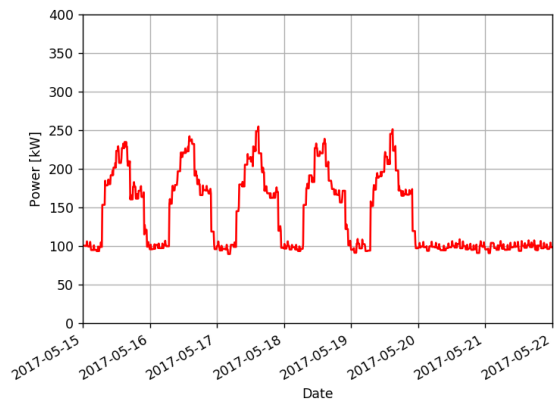
According to the data presented in the previous table, the average flow rate during the no-production period, which denotes the total leakage in the grid, represents around 16.5 % of the average flow rate during a normal operation period.

#### 7.3.2.2.2 NILM Estimation of Compressed Air Leakage – GreEn-ER Data with 10 Minutes Sampling Interval Study Case

In order to use NILM techniques to estimate the compressed air leakage in this case study, the same procedure used in the section 7.3.2.1.2 was then applied to this dataset. The following figures illustrate the load curves of the loads cited in section 7.3.2.1.2 and the global consumption, during the period used for the training and the period used to estimate the leaks. The following table presents the average power of these loads during both periods. The air compressor load curves were already shown in the previous section.

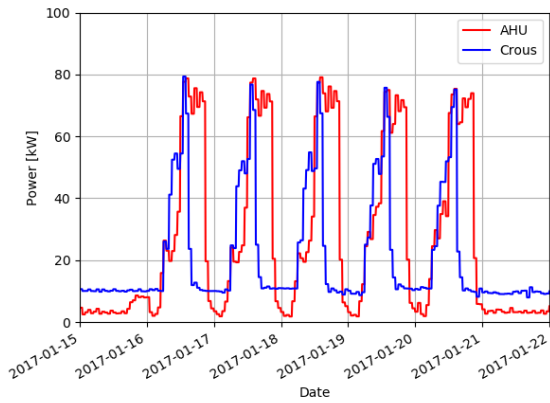


a) Training period

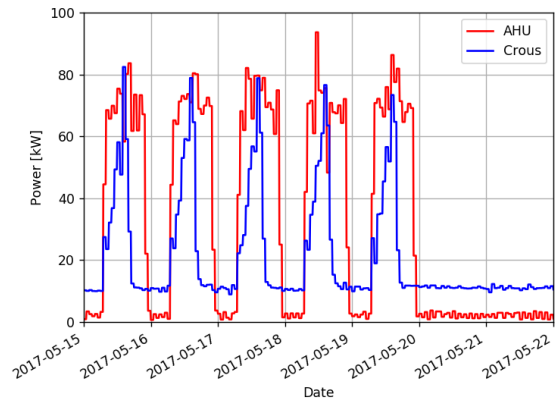


b) Test period

Figure 106 – Power data for the global consumption.

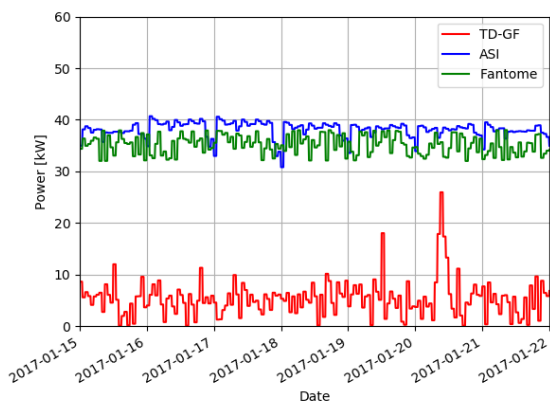


a) Training period

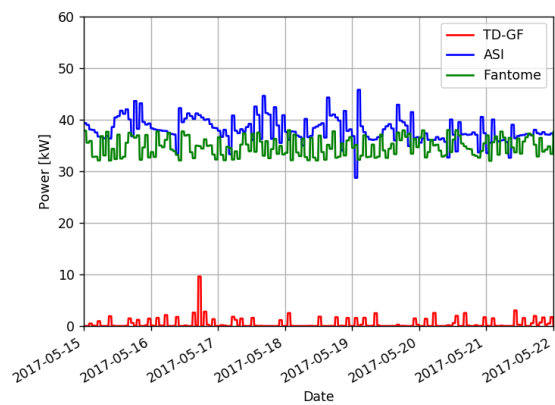


b) Test period

Figure 107 – Power data for the Crous and AHU loads.



a) Training period



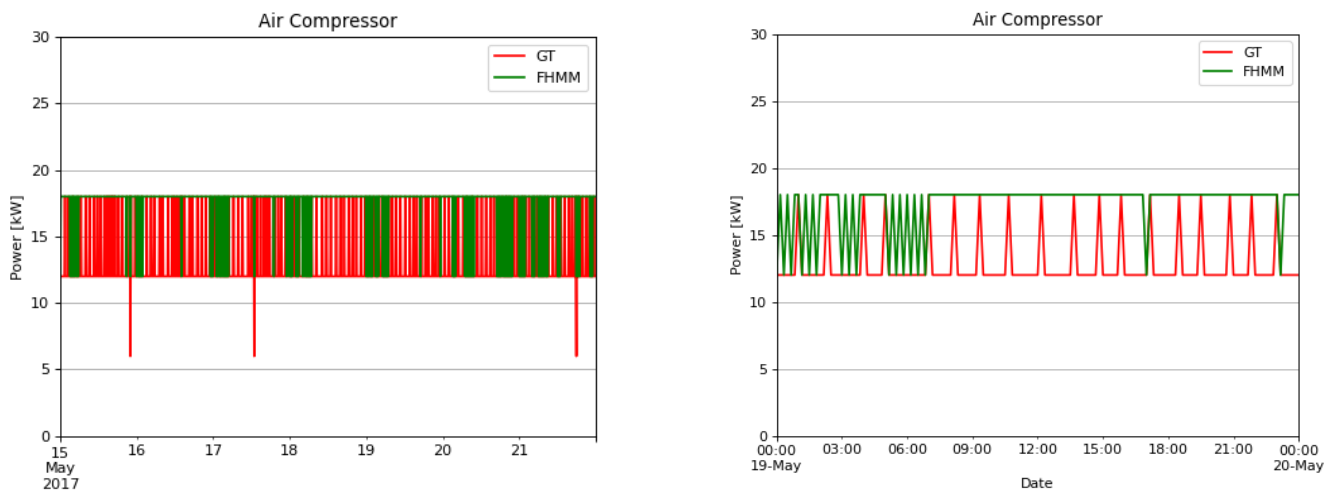
b) Test period

Figure 108 – Power data for the TD-GF, ASI and Fantome loads.

Table 40 – Average power of the loads.

Load	Average Power [kW]	
	Training Period	Test Period
Main	143.67	139.98
Crous	20.99	21.07
AHU	27.00	32.77
TD-GF	5.32	0.48
ASI	38.11	38.06
Fantome	35.18	34.78

With the help of the NILMTK, the FHMM algorithm is then applied, which results are illustrated in Figure 109, GT means Ground Truth, or the actual consumption, and FHMM are the results from the FHMM NILM algorithm.



a) Air compressor disaggregation results      b) Zoom in on the air compressor disaggregation results  
 Figure 109 – Disaggregation results.

It can be seen in the figure above that the FHMM algorithm was not able to retrieve most of the air compressor consumption pattern in this example. This can also be seen by the average power and flow rate estimation, calculated by applying the procedure described in the section 7.3.1. These results are exposed in Table 41.

Table 41 – Average air compressor power and flow rate during a no-production period.

	Average Power [kW]	Average Flow Rate [m <sup>3</sup> /h]
Estimation From Measurements	12.81	40.79
NILM Estimation	17.29	259.00

The results presented in Table 41 indicates that the NILM could not be really suitable to estimate the compressed air leaks in the context of an energy audit with this type of data.

However, the fact that the power data originally comes from cumulative energy measurements means that there can be an attenuation of power peaks, modifying the values of the powers associated with the states of certain equipment. To turn around this issue, power measurements were performed to the air compressor to well identify the power associated to both states. Figure 110 presents the data. In the same figure, it is highlighted the actual power associated with both load and unload states. One and a half day worth of data was measured with 5 minutes sampling interval.

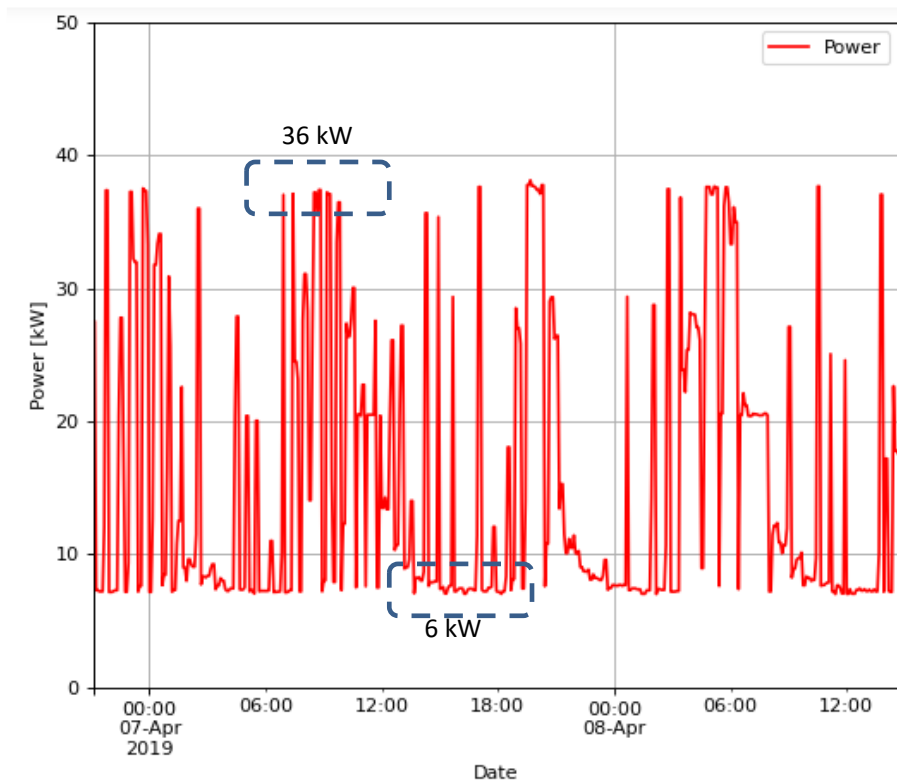


Figure 110 – Direct power measurement of the air compressor.

The fact that the two states (18 kW and 6 kW) identified in this dataset differ from the values associated to the air compressor due the 10 minutes sampling interval used, affect not only the flow rate estimation as also the performance of the FHMM. The 10 minutes interval attenuates the difference between the load and unload states. In this way, the importance of the value of the compressor load state, for example, in relation to the overall consumption is reduced. In cases like that, the disaggregation algorithms may also lose performance. Thus, the upsampling of the data, recovering information lost during the storage process in 10 minutes sampling interval could improve the results.

### 7.3.2.3 *Applying NILM Algorithms to Quantify Compressed air Leaks Using GreEn-ER Data with 1 Minute Sampling Interval*

As stated in the previous section, the data originally available concern the energy consumed, and are converted a posteriori into power. Therefore, the 10-minute sampling approach may hide some important features to the disaggregation task. Thus, seeking better estimations the data were upsampled into 1-minute sampling interval. The following sections exposes the compressed air leakage estimation, from measurements data and using the NILM algorithm FHMM, and the savings estimation with the repair of the leaks.

#### 7.3.2.3.1 Leakage Estimation from Measurements – GreEn-ER Data With 1 Minute Sampling Interval Study Case

In the previous section, the original load curve of the air compressor installed in the GreEn-ER building was exposed, in Figure 105. In that figure, the state with associated power of 18 kW was considered the load one while the unload state was associated with the 12 kW value. Usually, the power difference between these two states are more prominent than 6 kW. Actually, from direct power measurements, presented in Figure 110, the power associated with both load and unload states are 36 kW and 6 kW, respectively, which differs a little from the value presented in Table 30, probably taking in account also the dryer power. This can be explained by the 10-minutes sampling interval used to store the data, indicating that the air compressor changes its operating states at a greater rate than the sampling used. When measuring the energy consumed in the sampling interval, an average value is retrieved that reduces the difference between the powers associated to the two states. This fact may affect the flow rate estimation since it relies on how long the air compressor remains in the load and unload states, overestimating the flow rate during the periods in which the air compressor remains in the 18 kW state, and underestimating the flow rate during the 12 kW condition. Because of that, the upsampling to 1-minute sampling rate was performed.

In order to upsample the air compressor data keeping the original average power value it is necessary to determine how much time the compressor operates in each state. This is possible to estimate by performing a weighted average that can be translated to the linear system presented in the following equations.

$$P_{load}^{10} = \frac{(P_{load} * T_{load}^l) + (P_{unload} * T_{unload}^l)}{10} \quad \text{Equation 28}$$

$$P_{unload}^{10} = \frac{(P_{load} * T_{unload}^u) + (P_{unload} * T_{unload}^u)}{10} \tag{Equation 29}$$

$$10 = T_{load} + T_{unload} \tag{Equation 30}$$

In which,

$P_{load}^{10}$  and  $P_{unload}^{10}$  are the powers associated to both load and unload states in the 10 minutes sampling interval, equal to 18kW and 12kW, respectively,

$P_{load}$  and  $P_{unload}$  are the powers associated to both load and unload states in the 1 minute sampling interval, equal to 36kW and 6kW, respectively,,

$T_{load}$  and  $T_{unload}$  are the time that the compressor remains in load state during the 1 minute interval,

For each 10 minutes period that the compressor presents 18 kW of average power, it has operated 4 minutes in load condition (36 kW) and 6 minutes in unload condition (6 kW). For each 10 minutes period in which the air compressor power was 12 kW, the asset has operated 2 minutes in load conditions and 8 minutes in unload state. The load and unload operating conditions during the 10 minutes period were distributed in a random way. Figure 111 shows the air compressor power during the training and test periods, as well the two-hour moving average for better visualization. At the same time, Figure 112 presents a zoom in of those figures.

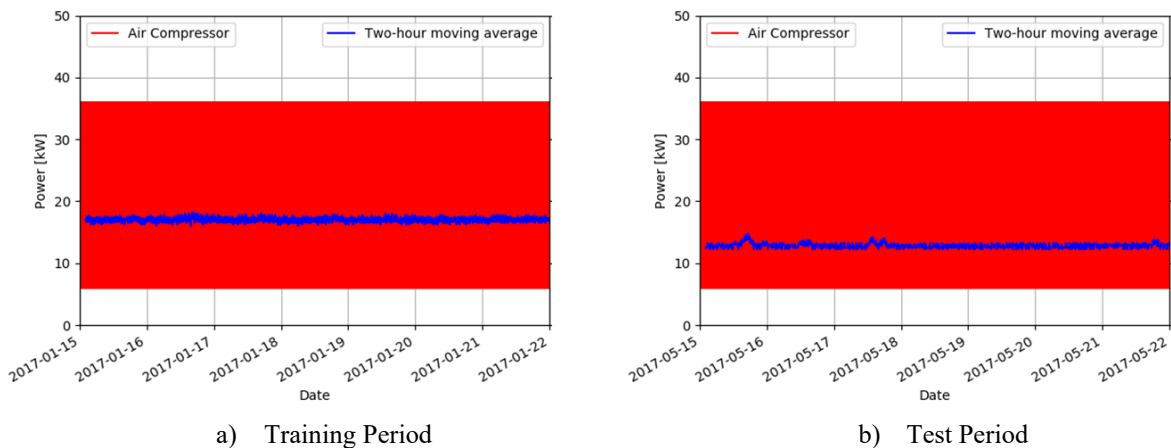
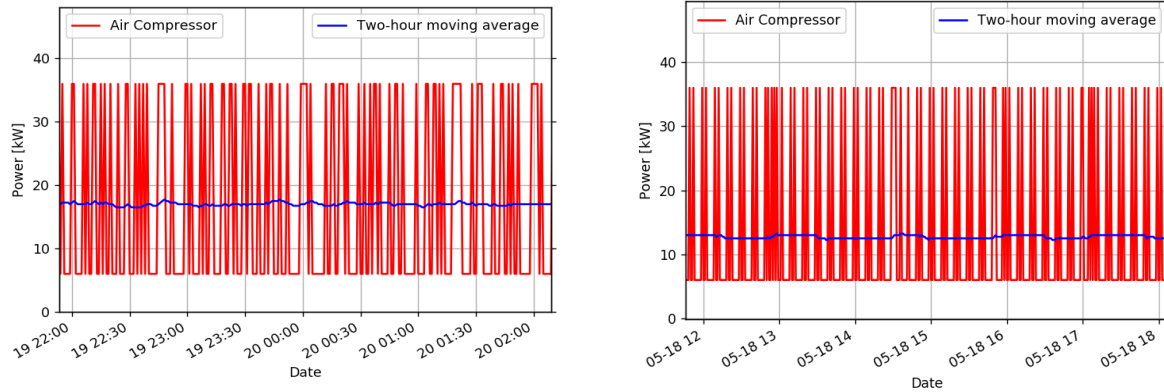


Figure 111 – Power data for the Air Compressor.



a) Training Period

b) Test Period

Figure 112 – Zoom in in the power data for the Air Compressor.

The upsampling of the data modifies the number of state changes in the operation of the air compressor. Consequently, the flow rate estimation drastically changes. With the 10-minutes sampling interval, the majority of the time during the training phase was in the load state. However, with the upsampling to 1-minute interval, one sample that was in load state in the 10 minutes sampling, belongs now only 40% of the time in this condition. So it is expected that the flow rate estimation decreases. This is confirmed by the values presented in Table 42. This table shows the number of samples in which each of the operating states of the air compressor were identified. In addition, using the procedure presented in section 7.3.1 the flow rate load curve was then estimated. The same table exposes the average power and estimated flow rate for the normal operation and no-production periods for both sampling intervals.

Table 42 – Average air compressor power and flow rate during the different operation periods

Sampling interval	Evaluated Period	Number of samples		Average value	
		Load	Unload	Power [kW]	Flow Rate [m <sup>3</sup> /h]
10 minutes	Training Period	846	163	17.03	246.51
	Test Period	140	869	12.83	40.79
1 minute	Training Period	3706	6375	17.03	108.08
	Test Period	2296	7785	12.83	66.96

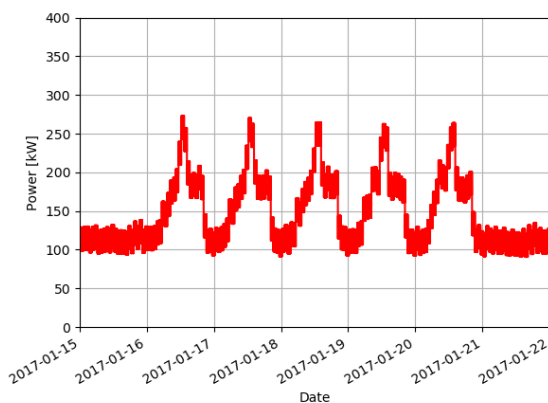
Considering the air compressor specifications presented in Table 30, the flow rate estimation presented in Table 42 for 1 minute sampling interval matches better to the air compressor characteristics when compared to the estimation for 10 minutes sampling interval. Seeing average power values, the air compressor operates at 52% of its rated capacity. In view

of the 10 minutes sampling interval, the flow rate estimation represented around 84% of the rated capacity, which is inconsistent with the power consumption, since this percentage should be lower than the 52% of the power, because of the zero flow rate delivered during the unload state. On the other hand, the flow rate estimation considering the 1 minute sampling interval represents 36.8% of the air compressor rated capacity, which is more consistent considering the energy consumption and the operation of this type of air compressor.

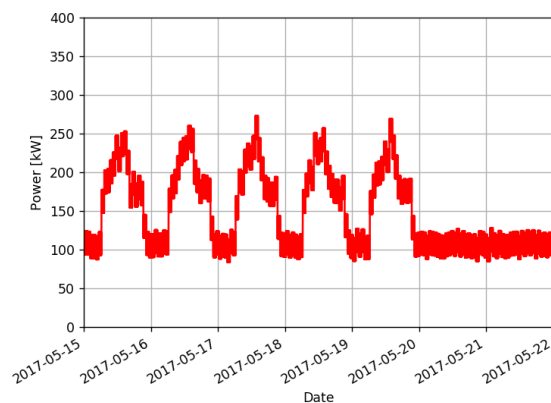
Assuming that the 1 minute sampling estimation is more accurate, the compressed air leakage represents 62.0% of the average flow rate during a normal operation period. However, 62% represents a high value of leakage, even in poor maintained environments and may suggest other problems, such as irrational usage of compressed air, non-optimal control and usage etc. Nevertheless, even if this estimation represents more than the leaks, it would suggest a more detailed investigation.

### 7.3.2.3.2 NILM Estimation of Compressed Air Leakage – 1 Minute Sampling Interval Study Case

In order to use NILM techniques to estimate the compressed air leakage in this case study, the same procedure used in the section 7.3.2.1.2 was then applied to this modified dataset. Since the sampling interval of the dataset originally 10 minutes, all these loads were upsampled to 1 minute sampling interval using forward fill, in order to match the sampling interval of the air compressor. The following figures illustrate the load curves of the loads cited in section 7.3.2.1.2 and the global consumption, during the period used for the training and the period used to estimate the leaks. The following table presents the average power of these loads during both periods. The air compressor load curves were already shown in section 7.3.2.3.1.



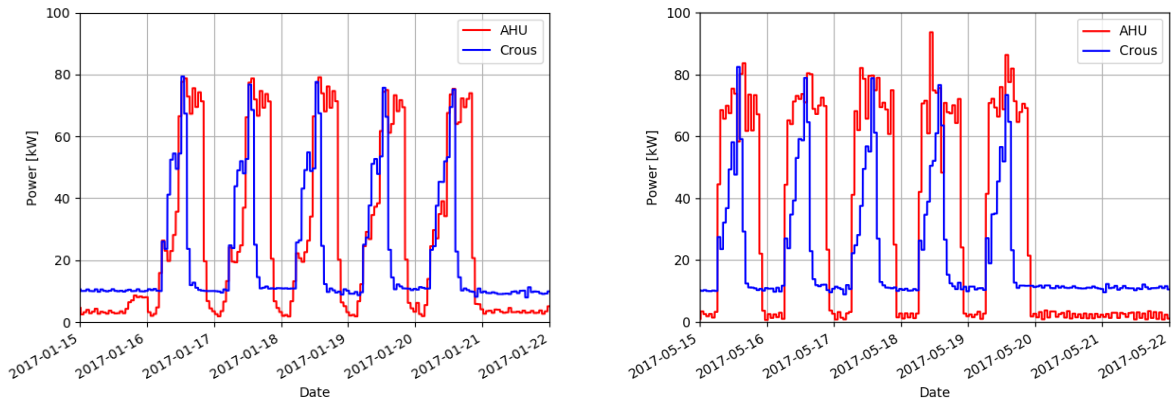
c) Training period



d) Test period



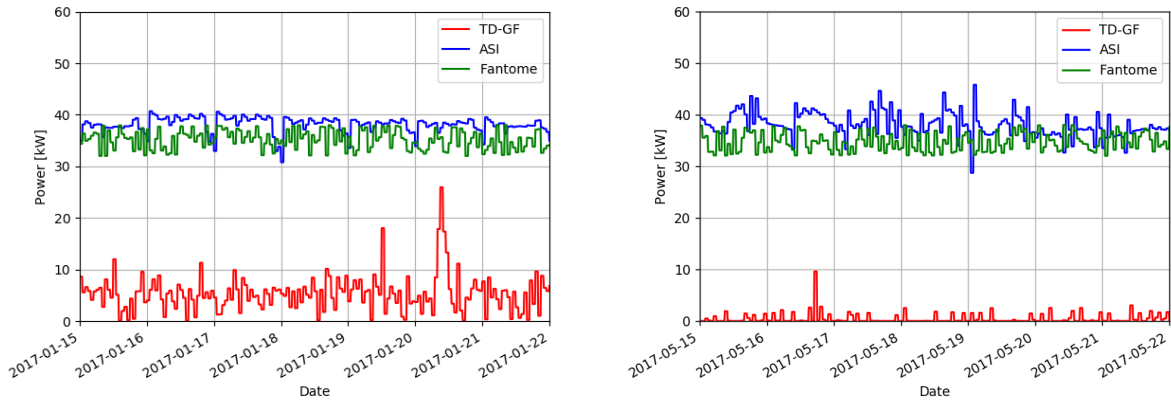
Figure 113 – Power data for the global consumption.



c) Training period

d) Test period

Figure 114 – Power data for the Crous and AHU loads.



c) Training period

d) Test period

Figure 115 – Power data for the TD-GF, ASI and Fantome loads.

Table 43 – Average power of the loads.

Load	Average Power [kW]	
	Training Period	Test Period
Main	143.67	139.98
Crous	20.99	21.07
AHU	27.00	32.77
TD-GF	5.32	0.48
ASI	38.11	38.06
Fantome	35.18	34.78

With the help of the NILMTK, the FHMM algorithm is then applied, which results are illustrated in Figure 116, GT means Ground Truth, or the actual consumption, and FHMM are the results from the FHMM NILM algorithm.

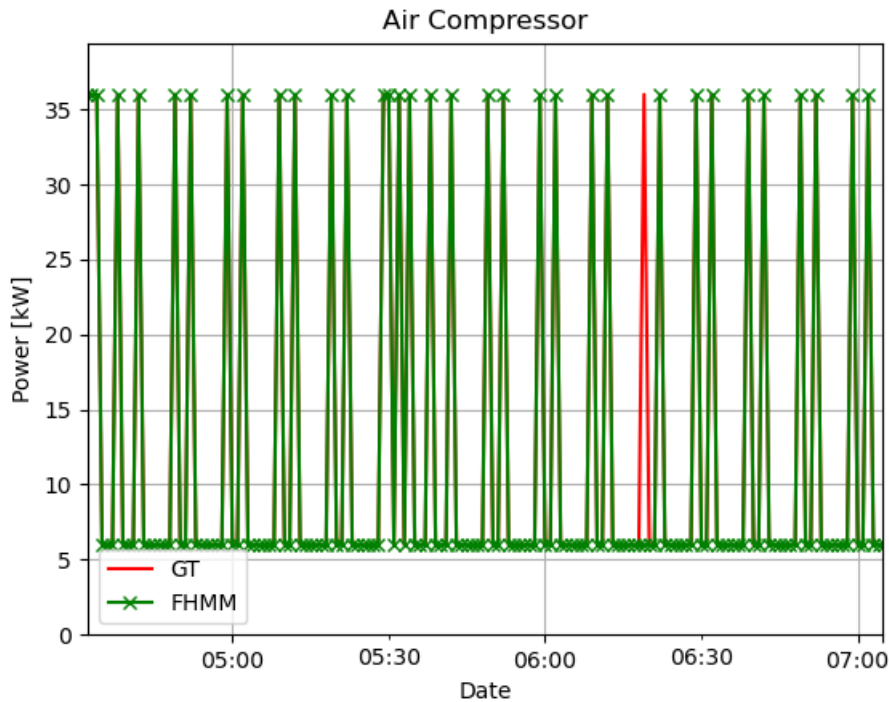


Figure 116 – Zoom in on disaggregation results.

It can be seen in the figure above that the FHMM algorithm was able to retrieve most of the air compressor consumption in this example. In order to quantify the estimation of the compressed air flow rate in this period, the association between flow rate and power showed in section 7.3.1 was applied. These results are exposed in Table 44. Then, the average compressed air flow was compared to the flow estimated of the actual compressed air power. These results are presented in Table 45.

Table 44 – Average air compressor power and flow rate during the different operation periods

Sampling interval	Evaluated Period	Number of samples		Average value	
		Load	Unload	Power [kW]	Flow Rate [m <sup>3</sup> /h]
1 minute	Training Period	3706	6375	17.03	108.08
	Test Period	2293	7787	12.82	66.88

Table 45 – Comparison between leakage estimation from measurements and from NILM.

	Average Power [kW]	Average Flow Rate [m <sup>3</sup> /h]
Estimation From Measurements	12.83	66.96
NILM Estimation	12.82	66.88

The results presented in Table 46 suggests that, it is possible to estimate the compressed air leakage using NILM techniques in the context of energy audits, considering the dataset investigated. With only one typical week used as training, the algorithm was able to estimate the compressed air flow with an error lower than 1% in this case.

Thus, the estimation presented in the last table could be used as a result of an energy audit in the compressed air system. These results allow the auditors to estimate the energy and financial savings with the repair of the leaks and, with the estimative of the costs of the repairs, calculate the payback time. Then, the client would have the necessary information to evaluate if the investment is worth or not.

### 7.3.2.3.3 Energy Savings With Leaks Repair – 1 Minute Sampling Interval Study Case

The leakage estimation performed in the previous section allows for the calculation of potential energy savings with its repair. Estimation of potential energy savings with leaks elimination is, therefore, presented in this section.

Although the complete elimination of compressed air leaks is theoretically feasible, it is nearly impossible to achieve in real environments. Because of that, potential savings were calculated considering incremental percentages repair of leaks, from 10% to 100%.

The determination of the compressed air flow rate after the leaks elimination allows the estimation of how long the air compressor would remain in both load and unload state and, hence, the average power of the air compressor. Table 46 and Table 47 present the average flow rate and power after the leaks repair in both periods, normal operation and no compressed air consumption.

Table 46 – Average air compressor power after the leaks repair during the normal operation period.

Leaks repaired [%]	Leaks repaired [m <sup>3</sup> /h]	Flow Rate [m <sup>3</sup> /h]	Number of samples		Average Power [kW]
			Load	Unload	
10	6.696	101.38	3476	6605	16.35
20	13.392	94.69	3247	6834	15.66
30	20.088	87.99	3017	7064	14.98
40	26.784	81.30	2788	7293	14.30
50	33.480	74.60	2558	7523	13.61
60	40.176	67.90	2328	7753	12.93
70	46.872	61.21	2099	7982	12.25
80	53.568	54.51	1869	8212	11.56
90	60.264	47.82	1640	8441	10.88
100	66.960	41.12	1410	8671	10.20

Table 47 – Average air compressor power after the leaks repair during the no compressed air consumption period.

Leaks repaired [%]	Leaks repaired [m <sup>3</sup> /h]	Flow Rate [m <sup>3</sup> /h]	Number of samples		Average Power [kW]
			Load	Unload	
10	6.696	60.26	2066	8015	12.15
20	13.392	53.57	1837	8244	11.47
30	20.088	46.87	1607	8474	10.78
40	26.784	40.18	1378	8703	10.10
50	33.480	33.48	1148	8933	9.42
60	40.176	26.78	918	9163	8.73
70	46.872	20.09	689	9392	8.05
80	53.568	13.39	459	9622	7.37
90	60.264	6.70	230	9851	6.68
100	66.960	0.00	0	10081	6.00

Considering the operation schedule, excluding vacations, holidays and weekends, the building operates 235 days per year in normal conditions. The remaining 130 days could be considered periods of no compressed air consumption. This information allows for the calculation of the average power, through a weighted mean calculation, and annual consumption by extrapolating these values to the whole year. The annual average power and energy consumption before the leaks repair are 15.53 kW and 136.1 MWh, respectively. Table 48 exposes the annual savings estimation with the compressed air leakage repair while Figure 116 illustrates the data.

Table 48 – Annual savings with the repair of compressed air leaks.

Leaks repaired [%]	Average Power [kW]			Annual consumption [kWh]	Annual savings [kWh]
	Normal operation	No compressed air consumption	Whole year		
10	16.35	12.15	14.85	130.1	5985.4
20	15.66	11.47	14.17	124.1	11970.8
30	14.98	10.78	13.48	118.1	17956.2
40	14.30	10.10	12.80	112.1	23941.6
50	13.61	9.42	12.12	106.2	29927.0
60	12.93	8.73	11.43	100.2	35912.4
70	12.25	8.05	10.75	94.2	41897.8
80	11.56	7.37	10.07	88.2	47883.2
90	10.88	6.68	9.38	82.2	53868.6
100	10.20	6.00	8.70	76.2	59854.0

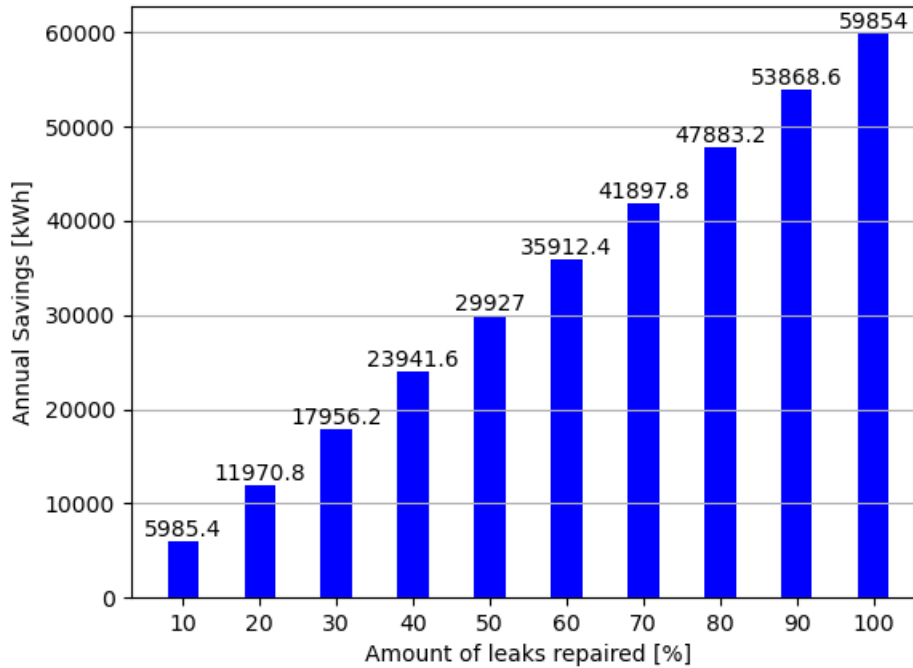


Figure 117 – Annual savings with the repair of compressed air leaks.

As an example, considering the ideal elimination of all leaks, savings of 59830.8 kWh represent around 44% in the compressed air system and 4.75% of the current annual global consumption. Nevertheless, it is important to remember that, although it is possible, the estimated leakage represents a high percentage compared to the total flow in the system, which may indicate other problems as irrational usage of compressed air, non-optimal control, and usage etc. However, even if this estimation represents more than the leaks, it leads itself to a more detailed investigation.

## 7.4 Conclusions

This chapter aimed to employ a NILM techniques to enhance energy audits by quantifying the leakage in a compressed air system of a tertiary building in a context similar to a real energy audit. One way to estimate the quantity of leakage in a compressed air system is from the association between the flow rate generated by the compressors and the power developed by them during non-production periods. During these periods, compressed air consumption should be zero, and all the air generated by the compressors is addressed to attend to the leaks. The quantification of compressed air leakage in a grid is a common analysis during an energy audit that usually yields good results in an energy audit. Even if an

energy audit can make use of both historical data collected from energy management systems and from measurements taken on site by the auditors, the period available for the measurements is limited to usually a few days or weeks, so it is possible that these measurements do not cover a non-production period, and therefore, making the estimation of the leakage in the grid tougher. In addition, it is unusual for some equipment, such as air compressors, to be monitored individually, while the measurement and storage of load curve data of global consumption is more common.

In a first step, in order to evaluate the possibility of using NILM techniques to quantify compressed air leaks in the context of an energy audit, an evaluation with synthetic data, created from a real compressor with the expected behavior in normal operation periods and with no compressed air consumption, was performed. The FHMM was elected to be used in this work because of its good performance when dealing with simple and moderate complex loads. However, the FHMM is not most suitable algorithm to tackle the NILM task in the presence of more complex loads, such as multi state appliances, or continuously variable loads [15]. This data has one hour sampling interval. The results obtained showed that it would be possible to use NILM algorithms to estimate the amount of leakage even with a short, like one week, training period. The difference between the leakage estimation calculated from the measured data and the results of applying the FHMM as NILM technique was less than 4%. Considering the calculations of annual energy savings from repairing the leaks, the difference in the estimation was even smaller, less than 1%.

With the promising results obtained from the synthetic data, the next step was to make the same evaluation but replacing the synthetic data with the actual data from an air compressor installed on the GreEn-ER. However, using original data from the dataset, with 10 minutes sampling interval, the flow rate was poorly estimated, making the leakage estimation also seem incorrect. With a deeper analysis of the data, it was identified that the power associated to the load and unload states are not correct. The fact that the power data originally comes from cumulative energy measurements means that there can be an attenuation of power peaks, modifying the values of the powers associated with the states of certain equipment, if the air compressor changes its states at a higher rate than the sampling interval, which seems to happen in this case. This problem may affect not only the flow rate estimation, as it relies on how long the equipment remains in load condition, but also the performance of the FHMM algorithm. The 10 minutes interval attenuates the difference between the load and unload states of the air compressor. In this way, the importance of the value of the compressor load

state, for example, in relation to the overall consumption is reduced. In cases like that, the disaggregation algorithms may also lose performance. Thus, the upsampling of the data, recovering information lost during the storage process in 10 minutes sampling interval could improve the results.

Furthermore, the upsampling to 1 minute sampling interval was chosen. It was made in a manner to maintain the same average power with the correct instant power associated to each state. The results obtained from the flow rate estimation seemed more accurate. With 1 minute sampling interval, the average power represents around 52% of the rated power while the flow rate estimation represents 36.8% of its rated value. At the same time, the same average power represents around 84% of rated flow rate with the 10 minutes sampling data, which is inconsistent. Assuming that the 1 minute sampling estimation is more accurate, the compressed air leakage represents 62.0% of the of the average flow rate during a normal operation period.. However, the high percentage of leaks estimated, around 62%, may suggest another problems on the system besides the leakage, irrational usage of compressed air or non-optimal control and operation of the assets. Anyway, considering that the estimation performed actually represents the leaks, the savings with the repair of all these leaks could reach about 44% in the compressed air system and 4.75% of the current yearly global consumption.

Considering the previous paragraphs it can be stated that it was possible to estimate the leaks using NILM techniques, in this case the FHMM, with some constraints, such as the importance of the difference between the power associated to both load and unload states with respect to the global consumption.

Although several studies were developed using the FHMM as NILM technique, it can be observed that this algorithm has good performance only for simple and moderate complex patterns, such as a fixed-speed air compressor. However, when more complex devices are on the mix of appliances, such as multi-state ones, continuously variable loads or even appliances with similar power in a state, the performance of the FHMM decays, and another algorithm may present better results.

## REFERENCES

- [1] Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC Text with EEA relevance OJ L 315, 14.11.2012, p. 1–56. <http://data.europa.eu/eli/dir/2012/27/oj>
- [2] Hart, G. W.. Nonintrusive appliance load monitoring. in *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870-1891, Dec. 1992, doi: 10.1109/5.192069.
- [3] Thumann, A., Niehus, T., Younger, W. J.. *Handbook of Energy Audits*, Ninth Edition (9th ed.). River Publishers 2012.
- [4] Benton, N.. Compressed Air Evaluation Protocol. In NREL – National Renewable Energy Laboratory. *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures* Subcontract Report NREL/SR-7A40-63210 November 2014  
<https://www.energy.gov/sites/prod/files/2015/01/fl9/UMPChapter22-compressed-air-evaluation.pdf>
- [5] Zhang, J.. Designing a cost-effective and reliable pipeline leak-detection system. *Pipes & pipelines international*, Issue 42, 1997, Pages 20-26.
- [6] Dudić S., Ignjatović, I., Šešlija, D., Blagojević, V., Stojiljković, M.. Leakage quantification of compressed air using ultrasound and infrared thermography, *Measurement*, Volume 45, Issue 7, 2012, Pages 1689-1694, ISSN 0263-2241, <https://doi.org/10.1016/j.measurement.2012.04.019>.
- [7] Eret, P., Meskell, C.. Microphone Arrays as a Leakage Detection Tool in Industrial Compressed Air Systems. *Advances in Acoustics and Vibration*, 2012, 1–10. <https://doi.org/10.1155/2012/689379>
- [8] Dudić S., Milenkovic, I., Šešlija, D., Blagojević, V., Stojiljković, M.. Leakage quantification of compressed air on pipes using thermovision. *Thermal Science*. Issue 16, 2012, Pages 571-581. Doi: 10.2298/TSCI120503191D.
- [9] Dindorf, R., Wos, P.. Test of measurement device for the estimation of leakage flow rate in pneumatic pipeline systems. *Measurement and Control* 51, 2018, Pages 514 - 527.
- [10] Eret, P., Harris, C., O'Donnell, G., Meskell, C.. A practical approach to investigating energy consumption of industrial compressed air systems. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*. 2012; 226(1):28-36. doi:10.1177/0957650911423173



- [11] Dindorf, R.. Estimating Potential Energy Savings in Compressed Air Systems. *Procedia Engineering*, 39, 204–211. 2012. <https://doi.org/10.1016/j.proeng.2012.07.026>
- [12] Piber, M. A.. Improving shipboard maintenance practices using non-intrusive load monitoring. Master's dissertation, 2007.
- [13] Berges, M. E., Goldman, E., Matthews, H. S., Soibelman, L.. Enhancing Electricity Audits in Residential Buildings with Non-intrusive Load Monitoring. *Journal of Industrial Ecology*, 14(5), 844–858. 2010 <https://doi.org/10.1111/j.1530-9290.2010.00280.x>
- [14] BELAiR. Centrale à vis lubrifiées fixe – ENCAP 30 datasheet 2020 <https://www.belair.fr/wp-content/uploads/2020/03/DA-FP-ENCAP30-8B-2020.pdf>
- [15] Gopinath, R., Kumar, M., Prakash Chandra Joshua, C., Srinivas, K.. Energy management using non-intrusive load monitoring techniques – State-of-the-art and future research directions, *Sustainable Cities and Society*, Volume 62, 2020, 102411, ISSN 2210-6707, <https://doi.org/10.1016/j.scs.2020.102411>.
- [16] Martin Nascimento, G. F., Delinchant, B., Wurtz, F., Kuo-Peng, P., Jhoé Batistela, N., and Laranjeira, T., “GreEn-ER - Electricity Consumption Data of a Tertiary Building”, *Mendeley Data*, V1, 2020 <http://dx.doi.org/10.17632/h8mmnthn5w.1>
- [17] Martin Nascimento, G. F., Delinchant, B., Wurtz, F., Kuo-Peng, P., Jhoé Batistela, N.. “Compressed Air Leakage NILM”, 2021 [Online]. Available: <https://gricad-gitlab.univ-grenoble-alpes.fr/martgust/compressed-air-leakage-nilm>
- [18] Martin Nascimento, G.F.; Wurtz, F.; Kuo-Peng, P.; Delinchant, B.; Jhoé Batistela, N. Outlier Detection in Buildings’ Power Consumption Data Using Forecast Error. *Energies* 2021, 14, 8325. <https://doi.org/10.3390/en14248325>
- [19] Atlas Copco. Compressor data sheet – GA18+-100 datasheet 2019 <https://www.atlascopco.com/content/dam/atlas-copco/local-countries/united-states/documents/cagi-data-sheets/ga-series-2020/ga11----30/GA18+-100%20-%20100%20psi%20-%20Air%20Cooled.pdf>

---

## 8 General Conclusions and Directions for Future Works

This chapter presents the general conclusion of this thesis, its major contributions and perspectives for related future works.

### **Chapter Contents**

<b>8.1 General Conclusions</b>	<b>186</b>
<b>8.2 Major Contributions</b>	<b>189</b>
<b>8.3 Future Directions</b>	<b>190</b>
<b>8.4 Publications</b>	<b>191</b>

## 8.1 General Conclusions

Buildings play a key role in the challenges of the energy transition. Worldwide, the energy consumed in buildings represents a significant part of the energy consumption. Considering electrical energy alone, in 2019 about 50% were consumed in residential or tertiary buildings, slightly above the industrial consumption share (42%). Thus, the importance of studies concerning energy efficiency and sobriety in buildings increased, leading to several research concerning these environments.

Data-driven approaches combined with machine learning techniques are more and more frequent in the energy sector, with the use of Artificial Intelligence (AI) and machine-learning methods. Because of that, the importance of the availability of datasets containing real measurement data of energy consumption is in constant increase. However, most of publicly available datasets concern residential buildings, and their typical loads. The availability of those datasets moved forward the research in the residential sector in some areas such as the non-intrusive load monitoring, leaving the tertiary sector behind. Hence, an open dataset, described in **Chapter 2**, containing the electricity consumption of the GreEn-ER building, containing both aggregated and disaggregated consumption in several levels was made available. It is hoped that this dataset will enable the advancement of research related to energy consumption in tertiary buildings as much as the availability of several residential dataset did for this sector.

Nevertheless, real data come along with real problems, translated into data quality issues. The increase of the use data-driven approaches raises a concern about the quality of the data used, as it is expected that poor data quality harm the performance of machine-learning algorithms, especially the supervised ones. Hence, **Chapter 3** made an overview about data quality problems and assessed the data quality of a subdataset of the GreEn-ER building in terms of accuracy and completeness. Moreover, that chapter showed how harmful data quality problems could be in machine-learning techniques performance. This is made through an example of the GreEn-ER energy consumption prediction using both healthy data and data with quality issues as training dataset.

In the light of what was discussed in Chapter 3, **Chapter 4** presented an algorithm, called Forecast Error, that combines machine learning techniques and classic statistical methods to detect outliers present within the GreEn-ER dataset. This method combined the random forest algorithm to predict the consumption using healthy data for training. A statistic

method, called Adjusted Boxplot is then applied to every sample of the error between the actual value and the prediction. This combination allowed the detection of most of the outliers present in both datasets tested, with an F-Score higher than 0.92, while the best statistic method alone reached 0.83. The Forecast Error method and the results obtained with its application to the GreEn-ER dataset were object of a journal publication.

The identification of potential energy efficiency measures rely on analyses of energy consumption data. These analyses are facilitated when consumption data to the appliance level are available. However, the availability of this type of data is not the standard for most of the facilities. To turn around this issue, research have focused on the development of non-intrusive load monitoring methods. This type of monitoring aims to extract individual energy consumption from the global aggregated one. Therefore, **Chapter 5** presents some of the most popular energy disaggregation algorithms, based both on the activation of finite states and artificial neural networks. It also presented the NILMTK, a framework developed by Batra et al designed to enable comparative analysis of different methods using several datasets, aiming to enhance the reproducibility of research in this field.

The algorithms presented in Chapter 5 were developed mostly based on residential datasets. However, the energy consumption in tertiary environments is typically different from the pattern found in residential buildings. Commercial buildings have many more appliances than a single household, which means that states' switch is more frequent, making it much more likely to exist more than one appliance changing its state simultaneously. In addition, tertiary buildings often have loads that, individually are small, but may represent an important share when combined with eventual similar loads. Laptop computers and chargers, among other electronic loads, are responsible for significant portion of tertiary building's electricity consumption, but when individually compared to the global consumption do not represent an important share. Small step changes are nearly impossible to identify and segregate from the global consumption. It can be hard to tell if these changes are caused by switching on a desktop computer or if it is due to larger continuously variable loads changing their demand. Because of that, it is unsure that these algorithms present satisfactory results when applied to tertiary buildings. Therefore, **Chapter 6** presented some insights about the energy disaggregation in tertiary buildings and how to overcome the limitations of NILM techniques in larger facilities, such as commercial buildings. The discussion indicated that it is crucial to elect target loads, decreasing the number of appliances and disregarding less

important loads, reducing the computation time and improving the performance of the techniques.

The same chapter applied the algorithms presented in Chapter 5 to a subset of the GreEn-ER dataset. Three major consumers, the compressed air system, the chillers and the restaurant, from the TGBT2 were chosen as target loads. The results indicated that, even with the reduction of the number of loads, the disaggregation in a commercial building is not trivial. Algorithms using artificial neural networks presented good results, but at a high computational cost, what can be a limitation especially in embedded or portable solutions. On the other hand, the FHMM was less precise, but with average performance, indicating that it could be used to target loads with simple patterns, such as fixed-speed rotary screw air compressors.

The path towards the reduction of energy consumption passes through the realization of energy audits. These analyses usually use on-site measured data by the auditors. However, the time available for these measurements is limited and may not include some modes of operation of certain appliances. One example of that is the quantification of compressed air leaks, which can be done by estimating the flow rate during a no compressed air consumption period. Nevertheless, these periods often do not coincide with auditors' schedule, issue that could be addressed by using historical data. However, historical data from energy management systems usually are only available for global consumption, and rarely for individual appliances. In this context, a NILM approach would be helpful to enhance energy audits carrying analysis of modes of operation not included in the on-site measurements.

Hence, **Chapter 7** comes as a liaison between energy disaggregation techniques and energy audits by estimating the compressed air leakage using a subdataset of the GreEn-ER dataset, using only one week of data as training set, against several months used by earlier applications. Three approaches were tested. Firstly, to assess the possibility of using NILM techniques to enhance energy audits, synthetic well-behaved data of a fixed-speed rotary screw air compressor were created from catalog data. Because of the nature of the load, the FHMM algorithm was elected to disaggregate the air compressor consumption from the global load. The results obtained confirmed the possibility.

With the positive results obtained from the synthetic data, it was time to test the hypothesis with real data from the GreEn-ER dataset. Hence, a subdataset containing data

from major consumers from both TGBT1 and TGBT2 was assembled. Using raw data, with 10 minutes sampling interval, the estimates obtained were not satisfactory.

Why does not the same approach work with real data? Taking a closer look at the data, the power values assumed to be associated with the load and unload states were not correct. Because of the nature of the data, originally cumulative energy, converted into power a posteriori, every sample represents the average power during the sampling interval. In the case when the air compressor alternates between both states at a faster rate than the sampling interval, which is common considering 10 minutes interval, the difference between the power associated to both states is attenuated. In this way, the importance of the value of the compressor load state, for example, in relation to the overall consumption is reduced. Thus, seeking better estimations the data were upsampled into 1-minute sampling interval. Considering the upsampled data, similar results to those obtained using synthetic data were achieved, indicating that the possibility of using NILM to enhance energy audits is real.

The fact that the FHMM presented good results indicates that there is no magical way to disaggregate typical loads from tertiary buildings and the choice of the algorithm used must be based on the target load consumption pattern.

## 8.2 Major Contributions

This thesis focused on the energy consumption of a tertiary building, called GreEn-ER and ways to reduce its consumption. To this, a dataset containing both aggregated and disaggregated data was made available, hoping that it can lead the advancement of research in data-driven approaches regarding this type of building.

However, the real data often come with data quality issues and the GreEn-ER dataset was no exception. In the light of these issues, a method to detect outliers, even the local ones, combining energy consumption prediction and a statistic technique, called adjusted boxplot, was developed. This method was object of a journal publication available in open-access.

The GreEn-ER building is a smart building, with more than 300 electricity meters, which allow for detailed consumption analyses. However, this is not the standard of the vast majority of tertiary buildings. Because of that, non-intrusive load monitoring shows potential to provide the information needed to identify prospective energy efficiency measures. However, most of the research on this field concerns residential buildings, which present

different consumption patterns. Considering this, this thesis presented some insights about the application of NILM techniques to data issued from a tertiary building. It was showed that there is not a “universal” algorithm suited for the tertiary building environment, but different algorithms, depending on the goal, can address each type of load.

Additionally, this thesis showed that it is possible to enhance energy audits by using energy disaggregation to quantify the compressed air leakage in the GreEn-ER building. The analyses presented in Chapter 7 were also object of a journal publication.

### 8.3 Future Directions

During the work performed in this thesis, some insights about future works arose:

- This work tried to address the lack of available datasets of energy consumption of tertiary buildings by publicizing the GreEn-ER dataset. Although the GreEn-ER building can be classified into the “education” category, it can be considered a quite complete building, containing a restaurant, HVAC systems, several office loads, among others. However, the tertiary sector is very diverse, with different types of activities, which lead to different energy consumption patterns, like in the case of hospitals, for instance. Hence, the gap is not yet filled when compared to the residential sector. The continuous publication of dataset regarding energy consumption in tertiary buildings, or even in industrial facilities, is of great importance for the advancement of the research on the application of machine-learning and artificial intelligence techniques on these environments.
- This work showed that the data quality of the measures performed in the GreEn-ER building could be highly improved. A real-time application to monitor the data quality would enhance the maintenance of the system reducing costs and information loss.
  - One example is the outlier detection algorithm presented in this thesis that looks for anomalies within historical data. One step further would be implementing such algorithm in an real time application, once a good prediction model is defined using healthy data.
- During this work, a subset of the GreEn-ER dataset was integrated to the NILMTK. One step further into the advancement of NILM research in

environments outside the residential sector would be integration of all the GreEn-ER dataset to the framework, opening possibilities to test new or old algorithms in several types of loads.

- The quantification of compressed air leakage using NILM techniques when more than one compressor is needed to fulfill the leaks, or a variable-speed compressor is operating is yet to be addressed.
- The functioning of a fixed-speed rotary screw air compressor is, usually, independent from the environment where it is installed. Therefore, the possibility of NILM algorithms to profit of a pre-existent library of power signatures of such appliances in order to disaggregate the air compressor consumption even without measurements on-site can be a promising research direction.

## 8.4 Publications

The work performed on the present thesis allowed the publication of a dataset, the co-authorship in a conference article, along two journal papers. These works are listed as follows:

- **Dataset:**  
 Martin Nascimento, G. F.; Delinchant, B.; Wurtz, F.; Kuo-Peng, P.; Jhoe Batistela, N.; Laranjeira, T. GreEn-ER - Electricity Consumption Data of a Tertiary Building. *Mendeley Data*, V1, 2020  
<http://dx.doi.org/10.17632/h8mmnthn5w.1>
- **Conference Article:**  
 Cesário Pereira Pinto, J. O., Martin Nascimento, G. F., Wurtz, F., Delinchant, B., Moreto, M., Kuo-Peng, P.; Jhoe Batistela, N.. Architecture Multi-Agents pour Surveiller la Qualité des Données de Consommation d'Énergie. *IBPSA France 2020*, 2020, Reims, France. (hal-03345308)
- **Journal Articles:**  
 Martin Nascimento, G.F.; Wurtz, F.; Kuo-Peng, P.; Delinchant, B.; Jhoe Batistela, N. Outlier Detection in Buildings' Power Consumption Data Using Forecast Error. *Energies* 2021, 14, 8325. <https://doi.org/10.3390/en14248325>

Martin Nascimento, G.F.; Wurtz, F.; Kuo-Peng, P.; Delinchant, B.; Jhoe



---

Batistela, N. Quantifying Compressed Air Leakage through Non-Intrusive Load Monitoring Techniques in the Context of Energy Audits. *Energies* 2022, 15, 3213. <https://doi.org/10.3390/en15093213>

## Résumé de la thèse

### Introduction

Les bâtiments, qu'ils soient résidentiels ou tertiaires, jouent un rôle clé dans la transition énergétique car ils représentent une partie importante de la consommation énergétique mondiale et des émissions de gaz à effet de serre. Par conséquent, la réalisation d'audits énergétiques par une équipe spécialisée dans de tels environnements peut permettre de réaliser d'importantes économies d'énergie. Ces audits sont des outils puissants qui permettent d'identifier et de quantifier le potentiel d'économies dans ces environnements.

Cependant, en raison du temps limité dont dispose l'équipe pour collecter les données des mesures sur site, ou du manque de données historiques, on peut dire que les audits énergétiques prennent un instantané des conditions d'une installation à ce moment précis. Ainsi, certains modes de fonctionnement, qui peuvent cacher des potentiels d'économies, peuvent ne pas être abordés par les analyses. Par conséquent, l'utilisation de méthodes NILM (Non Intrusive Load Monitoring) peut être utile pour améliorer les analyses effectuées, en récupérant la courbe de charge des modes de fonctionnement non monitorés, augmentant potentiellement les économies d'énergie identifiées. Ainsi, cette thèse cherche à établir un lien entre les analyses effectuées dans le cadre d'un audit énergétique et l'utilisation des méthodes NILM comme moyen d'améliorer ces analyses.

Toutefois, lorsqu'on énumère les études déjà réalisées sur la consommation d'énergie dans les bâtiments, on constate que la plupart se sont concentrées dans le secteur résidentiel. Cela peut s'expliquer par plusieurs facteurs, de la présence de charges plus simples et moins nombreuses dans les bâtiments résidentiels à la disponibilité de plusieurs jeux de données contenant des données sur la consommation d'énergie de ce type de bâtiment. En revanche, il n'y a pas beaucoup de jeux de données qui concernent les bâtiments tertiaires. De plus, les charges typiques des bâtiments tertiaires sont plus complexes et nombreuses par rapport au secteur résidentiel. Alors, cette thèse vise également à augmenter la disponibilité des données de consommation d'énergie en publiant des jeux de données contenant les données du bâtiment GreEn-ER, situé à Grenoble, en France.

Pourtant, les données de consommation réelles entraînent, intrinsèquement, des problèmes de qualité, en raison à des défaillances d'acquisition ou de stockage. Par conséquent, cette thèse propose également d'analyser et de quantifier certains des problèmes de qualité des données présents dans les jeux de données GreEn-ER.

### Objectifs

Cette thèse a pour objectif de :

- Augmenter la disponibilité, sous une forme ouverte, des données de consommation d'énergie des bâtiments tertiaires. Pour ce faire, les jeux de données contenant les données de plus de 300 compteurs électriques du bâtiment GreEn ER est publié.
- Développer des méthodes, des modèles et des outils pour quantifier et analyser les problèmes de qualité des données de consommation d'énergie.
- Développer des méthodes, modèles et outils pour améliorer les audits énergétiques des bâtiments tertiaires.

•

### Méthodologie

Les analyses effectuées pour la préparation de cette thèse passent par l'analyse approfondie des données du bâtiment GreEn-ER, un bâtiment massivement surveillé contenant plus de 1500 capteurs, dont plus de 300 sont des compteurs électriques.

Ainsi, la première étape consiste à collecter, organiser et publier les données de consommation d'énergie électrique de ce bâtiment.

Dans un deuxième temps, il s'agit de quantifier et d'analyser la qualité des données collectées et de développer un algorithme pour détecter les éventuelles outliers présentes dans les jeux des données.

Le test des algorithmes de désagrégation énergétique existants, développés pour les environnements résidentiels, dans le contexte des bâtiments tertiaires constitue la quatrième étape de l'élaboration de cette thèse.

Pour finir, une relation entre la désagrégation énergétique et les diagnostics énergétiques est présentée, en quantifiant les fuites d'air comprimé existantes dans GreEn-ER.

## Résultats et Analyse

Le grand défi actuel est de lutter contre le changement climatique causé par les activités humaines, notamment celles liées à la consommation d'énergie, ce qui conduit à d'importantes discussions sur la transition énergétique. Dans ce contexte, les bâtiments jouent un rôle central, car, dans leur ensemble, ils comptent parmi les plus gros consommateurs d'énergie de la planète. Dans ce contexte, les bâtiments jouent un rôle central, car, dans leur ensemble, ils comptent parmi les plus gros consommateurs d'énergie de la planète. En ne considérant que l'énergie électrique, en 2019, environ 50% a été consommée dans les bâtiments résidentiels ou tertiaires, légèrement au-dessus de la consommation industrielle (42%). La forte consommation d'énergie indique également un grand potentiel d'économies dans ces environnements. Pour ces raisons, cette thèse se concentre sur l'étude des réductions potentielles de la consommation d'énergie dans les bâtiments, en particulier les bâtiments tertiaires, car les études existantes centrées sur les environnements résidentiels sont nombreuses.

Il y a plusieurs raisons pour lesquelles l'environnement résidentiel est plus étudié. Les charges typiques des bâtiments résidentiels sont plus simples et moins nombreuses, par rapport à celles des bâtiments tertiaires. En outre, la disponibilité des données de consommation concernant les environnements résidentiels est beaucoup plus grande. Donc, afin de remplir le vide dans la disponibilité des données des bâtiments tertiaires, un jeu de données a été rassemblé et organisé de manière à être exploré par la communauté scientifique. Ce jeu de données contient les données de consommation des compteurs d'électricité GreEn-ER. Il y a plus de 300 points de mesure qui correspondent à la fois aux mesures générales du tableau de distribution, dites agrégées, et aux charges spécifiques, dites désagrégées. Les données sont disponibles dans des fichiers au format CSV, et les métadonnées sont organisées en dessins et tableaux accessibles par des Notebooks qui combinent ces informations avec le langage de programmation Python.

Néanmoins, lors de l'analyse des données de consommation du bâtiment, il a été vérifié l'existence de plusieurs incohérences dans les données, traduites sous la forme de problèmes de qualité. Parmi les problèmes constatés, on souligne l'incomplétude des données, avec des échantillons manquants, et des problèmes de précision, qui se manifestent par la présence de valeurs aberrantes et d'incohérences lors de la comparaison de la somme des compteurs en aval par rapport au compteur en amont. Ces problèmes, lorsqu'ils sont très prononcés, peuvent entraîner des difficultés lors de l'utilisation de techniques d'intelligence artificielle pour effectuer certaines analyses. Par conséquent, la détection et la correction éventuelle de ces problèmes est une nouvelle voie de recherche qui s'ouvre en fonction des données utilisées.

L'un des problèmes le plus fréquent est la présence d'outliers, c'est-à-dire d'échantillons très différents des autres ou de ce qui est attendu. Il existe plusieurs techniques pour détecter ce type d'échantillon. Parmi les plus utilisées, on trouve deux méthodes statistiques simples qui utilisent des variables telles que la moyenne, l'écart-type ou même la médiane et l'écart interquartile, les plus populaires étant 3-Sigma et Boxplot. Il existe également des variantes de ces techniques pour tenir compte d'éventuelles asymétries dans la distribution des données,

comme le Skewed Boxplot ou le Adjusted Boxplot. Ces méthodes utilisent des indices, dont le plus simple est la valeur de l'échantillon elle-même. Toutefois, ces méthodes n'ont pas donné de bons résultats lorsqu'elles ont été appliquées directement aux données de consommation de GreEn-ER. Une nouvelle stratégie a donc été élaborée, en utilisant un autre indice. Puisqu'un outlier peut être considéré comme un échantillon qui s'écarte beaucoup d'une valeur attendue, un indice possible à utiliser est l'erreur de prédiction, ou la différence entre la valeur de l'échantillon et la valeur prédite. Pour cela, il est nécessaire que la prédiction effectuée soit fiable. Dans le cas de GreEn-ER, la prédiction de la consommation a été effectuée en utilisant comme méthode de prédiction le Random Forest, en choisissant manuellement des périodes avec des données saines comme jeu de données pour l'entraînement de l'algorithme. Pour évaluer les performances des techniques appliquées, le F-score a été utilisé, une métrique qui utilise le nombre de faux positifs, de faux négatifs et de vrais positifs. Parmi les techniques testées pour la détection des valeurs aberrantes, celle qui a obtenu le meilleur résultat, présentant le F-score le plus élevé, a été l'utilisation du Adjusted Boxplot appliqué à l'erreur de prévision, ce qui a été appelée méthode de l'erreur de prévision.

Indice	Méthodes	Outliers potentiels détectés	Vrais positifs	Faux négatifs	Faux positifs	Mauvaises classifications	F-Score
Valeur de l'échantillon	3 Sigma	6	6	206	0	206	0,055
	<i>Boxplot</i>	6	6	206	0	206	0,055
	<i>Adjusted boxplot</i>	271	172	40	99	139	0,712
Erreur de prédiction	3 Sigma	6	6	206	0	206	0,055
	<i>Boxplot</i>	860	212	0	648	648	0,396
	<i>Adjusted boxplot</i>	205	192	20	13	33	0,921

La consommation d'énergie dans les bâtiments est directement liée au comportement des habitants. Cependant, ce n'est pas facile pour quelqu'un qui n'est pas un expert de quantifier l'impact de son comportement sur la consommation d'énergie. Pour cette raison, plus le niveau de détail des informations sur la consommation est élevé, plus les chances d'identifier les économies d'énergie potentielles sont grandes. Ainsi, les informations sur la consommation en temps réel et avec les détails des équipements individuellement sont celles qui offrent le plus grand potentiel d'économies. La façon la plus directe d'obtenir ces informations est d'installer des compteurs individuels pour chaque appareil. Cependant, cette alternative est souvent peu pratique, notamment dans les bâtiments tertiaires, avec de nombreuses charges, ou dans les bâtiments anciens. Ainsi, une façon de contourner cette limitation est l'utilisation de méthodes de désagrégation d'énergie, qui utilisent des techniques d'intelligence artificielle pour récupérer la consommation individuelle des équipements à partir de la consommation globale du foyer. Il existe plusieurs méthodes développées à cet effet, parmi lesquelles on peut citer le *Combinatorial Optimization*, le *Factorial Hidden Markov Model* (FHMM), ou encore des méthodes basées sur des réseaux de neurones artificiels, comme le *Sequence-to-point*. Cependant, ces méthodes ont été développées sur la base de jeux de données provenant du secteur résidentiel, de sorte que le succès de leur application dans des environnements tertiaires ou même industriels est incertain.

Ainsi, certains de ces algorithmes ont été testés sur le jeu de données GreEn ER, également issu de ce travail. En raison du grand nombre de charges présentes dans le bâtiment, plus de 300 comme déjà mentionné, la réduction du problème, avec le choix des charges cibles n'est pas seulement recommandée, mais primordiale pour le succès de l'application des méthodes choisies. Parmi les algorithmes testés, celui qui a obtenu les meilleures performances globales

c'est le Sequence-to-point. Cependant, pour des charges simples, comme un compresseur d'air tout-ou-rien, par exemple, le FHMM a présenté des résultats satisfaisants.

Un premier pas vers la réduction de la consommation d'énergie consiste à effectuer un audit énergétique, qui consiste en une série d'analyses visant à quantifier la consommation et les pertes d'énergie grâce à une étude détaillée des performances des différents systèmes. Ces analyses comprennent des mesures sur site et la collecte de données historiques. Ensuite, des calculs spécifiques à chaque système sont effectués afin de déterminer ses performances et d'identifier les économies potentielles d'énergie et financières.

Une façon de procéder à la collecte de données consiste à intégrer l'équipe dans l'environnement à évaluer, ce qui lui permet de vivre le fonctionnement quotidien de l'établissement. Cependant, le temps dont disposent les auditeurs pour effectuer cette tâche est limité, généralement quelques jours ou semaines. Pour cette raison, certains modes de fonctionnement de certains équipements peuvent ne pas être mesurés pendant la période de collecte de données des auditeurs. En plus, il est rare de disposer de données historiques et de mesures individuelles des appareils alors que la consommation globale est généralement historisée. Ainsi, une façon de contourner ce problème est d'appliquer des méthodes de désagrégation énergétique pour récupérer les données de consommation des équipements cibles pendant les périodes où l'équipe qui effectue l'audit n'est pas sur place. Cependant, le succès de cette application est incertain, car les données pour l'entraînement des algorithmes sont limitées aux données mesurées par l'équipe, alors que certains mois sont utilisés dans la littérature à cette fin.

Pour tester la possibilité d'utiliser des méthodes de désagrégation énergétique afin d'améliorer les analyses effectuées dans le cadre d'un audit énergétique, on a cherché à quantifier les fuites d'air comprimé de GreEN-ER. Comme la principale charge cible à désagréger était un compresseur d'air à vis à vitesse fixe, dont le profil de fonctionnement est relativement simple, le FHMM a été choisi comme méthode de désagrégation. À cette fin, une semaine typique de fonctionnement du système a été choisie comme période d'entraînement pour les algorithmes. Une période pendant laquelle il n'y a pas eu de consommation d'air comprimé, pendant laquelle le profil de fonctionnement était différent de la période d'entraînement, a été utilisée pour la quantification des fuites. Les résultats obtenus sont présentés dans le tableau suivant.

	<b>Puissance moyenne [kW]</b>	<b>Débit d'air moyen [m<sup>3</sup>/h]</b>
Mesures	12,83	66,96
Désagrégation (NILM)	12,82	66,88

Les résultats obtenus suggèrent qu'il est effectivement possible d'estimer les fuites d'air comprimé en utilisant des techniques de désagrégation d'énergie dans le cadre d'un audit énergétique, compte tenu des données utilisées. Avec une seule semaine typique utilisée comme entraînement, l'algorithme a été capable d'estimer le débit d'air comprimé avec une erreur inférieure à 1%, dans ce cas. Ainsi, cette estimation pourrait être utilisée comme résultat d'un audit énergétique du système d'air comprimé, permettant d'estimer les économies énergétiques et financières avec la réparation des fuites et, avec le coût des réparations, de calculer le temps de retour de l'investissement. Ainsi, le client dispose des informations nécessaires pour effectuer, ou non, les travaux nécessaires.

## Conclusions

L'utilisation de l'intelligence artificielle est de plus en plus fréquente dans le secteur de l'énergie. C'est pourquoi il est de plus en plus important de disposer de jeux de données

contenant des données réelles issues de mesures, afin de pouvoir développer et tester des algorithmes. Cependant, la plupart des jeux de données disponibles en libre accès concernent les bâtiments résidentiels, compte tenu de leurs charges typiques. Pour cette raison, cette thèse a cherché à réduire cette lacune avec la publication des jeux de données GreEn-ER, un bâtiment tertiaire avec plus de 300 compteurs électriques. On s'attend à ce que cet ensemble de données permette l'avancement de la recherche liée à la consommation d'énergie dans les bâtiments tertiaires, comme la disponibilité de plusieurs ensembles de données résidentielles faites pour ce secteur.

En revanche, les données réelles s'accompagnent de problèmes réels, qui se traduisent par des problèmes de qualité des données. Le recours accru aux approches fondées sur les données soulève des inquiétudes quant à la qualité des données utilisées, car on s'attend à ce qu'une mauvaise qualité des données nuise aux performances des algorithmes d'apprentissage automatique. Ainsi, cette thèse a présenté un algorithme, appelé Forecast Error, qui combine des techniques d'apprentissage automatique et des méthodes statistiques classiques pour détecter les outliers présentes dans les jeux des données GreEn-ER. Cette méthode utilise l'algorithme Random Forest pour prédire la consommation en utilisant des données saines pour l'entraînement. Une méthode statistique, appelée Adjusted Boxplot, est ensuite appliquée à chaque échantillon de l'erreur entre la valeur réelle et la prédiction. Cette combinaison a permis de détecter la plupart des valeurs aberrantes présentes dans les deux jeux de données testés, avec un F-Score plus élevé.

L'identification des potentiels d'efficacité énergétique dépend des analyses des données de consommation d'énergie. Ces analyses sont facilitées lorsque les données de consommation des appareils individuels sont disponibles. Cependant, la disponibilité de ces données n'est pas habituelle pour la plupart des installations. Pour contourner ce problème, des recherches ont été menées en vue de développer des méthodes non intrusives de monitoring des charges visant à extraire la consommation d'énergie individuelle de la consommation globale. Ainsi, cette thèse a testé plusieurs algorithmes déjà existants dans la littérature, développés en considérant un environnement résidentiel, dans un bâtiment tertiaire, obtenant des résultats mixtes. En raison de la grande quantité de charges existantes, condition typique d'un grand bâtiment tertiaire, la réduction du problème, avec le choix des charges cibles, est de la plus haute importance pour le succès de l'application des algorithmes NILM.

Enfin, cette thèse a également cherché à établir un lien entre les techniques NILM et les diagnostics énergétiques en estimant les fuites d'air comprimé sur la base d'un sous-jeu du jeu de données GreEn-ER. Les résultats obtenus ont montré qu'il est possible d'améliorer les analyses effectuées dans le cadre d'un diagnostic énergétique en utilisant des méthodes d'intelligence artificielle pour récupérer les courbes de charge des équipements à partir de la courbe de consommation globale d'une unité.

**Mots-clés:** Bâtiments tertiaires, sobriété énergétique, efficacité énergétique, désagrégation énergétique, qualité des données, détection des outliers.