UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Aprendizado de Máquina
Eficiência energética dos edifícios

Aluno: Guillaume Jeusel

Professor: Alexandre G. Evsukoff

Disciplina: Inteligência Computacional

9 de novembro de 2016

Sumário

1 Introdução

1.1 Problema

Com uma demanda de energia sempre crescente nosso mundo, o problema de economia de energia é colocado no centro das preocupações. O conceito de *négaWatt* [1] se base sobre a ideá que é mais barato economizar energia do que comprar-lho. E um campo cujo desperdício de energia continua a ser importante é o edifício.

Por conseguinte, as investigações na área do desempenho energético dos edifícios cresceu muito recentemente; uma acção prioritária que as sociedades deve ter em mente é a redução do consumo de energia dos novos edifícios, também como a renovação dos antigos. A propósito, a legislação sobre o desempenho energético dos edifícios é sempre mas exigente, especificamente nos países europeus com a directiva 2002/91/CE limitando o consumo de energia dos edifícios [2].

1.2 Conhecimento Prévio

Para o design desses edifícios, é necessário a computação dos termos chamados *"Heat Load" e "Cooling Load"* (que pode ser traduzido pelo "carga de aquecimento" e "carga de arrefecimento" respetivamente). Eles são diretamente ligados à especificação dos equipamentos responsáveis para manter uma temperatura confortável, e então ao consumo energético. Esses coeficientes são dependentes das características geométricas dos edifícios, como também do clima local e do uso deles (industrial, casal ...).

Existem muitos diferentes software de simulação que são eficientes para prever o consumo energético dos edifícios em projeto com uma precisão aceitável. Eles resolvam as equações diferencias da termodinmica aplicada a uma geometria particular. No enquanto, essas simulações podem demorar muito tempo, sem mencionar que quando um parmetro é mudado, a simulação deve ser reinicializada desde ao início.

Desse fato, um interesso crescente sobre o uso das técnicas de aprendizado de máquinas nasci. A ideá é a seguinte: suponho que você tem um banco de dados recente com as características e cargas de um grande número de edifícios, o uso de estatísticas e aprendizado de máquinas pode reduzir o tempo de computação e facilitar o experimento de diversos parmetros. Nos podemos pensar até criar um banco de dados com os diferentes resultados de simulação, e depois prever o desempenho energético de um novo edifício com interpolação dos resultados que nos já temos.

Isto foi a ideá do engenheiro civil *Angeliki Xifara* e do matemático *Athanasios Tsanas* da universidade de Oxford. Usando o software Ecotect, um conjunto de dado foi criado da simulação do desempenho energético para 768 geometrias de edifícios, assumindo uma localização em Atena, Grécia e um uso residencial com sete pessoas. Nos vamos estudar esse banco de dados.

Para ter mais informações sobre as hipóteses de simulação, deve-se referir ao papel deles [3].

2 Descrição dos dados

2.1 Dados

O dataset é tirado do web-site UCI – Machine Learning Repository [4]. A figura 2.1 contem um resumo geral desse conjunto de dados.

Data Set Characteristics:	Multivariate	**Number of Instances:**	768	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	2012-11-30
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	95751

Figura 2.1: Características dos dados

Ele é composto de 768 registros e tem 8 variáveis de entrada e 2 de saídas que são as seguintes:

Tabela 2.1: Mathematical representation of the input and output variables

Mathematical representation	Input or output variable	Number of possible values
X1	Relative Compactness	12
X2	Surface Area	12
X3	Wall Area	7
X4	Roof Area	4
X5	Overall Height	2
X6	Orientation	4
X7	Glazing Area	4
X8	Glazing Area Distribution	6
y1	Heating Load	586
y2	Cooling Load	636

É importante de notar que as variáveis de entradas são descontinuidades, e que não tem valores ausentes.

2.2 Estatísticas do conjunto de dados

2.2.1 Variáveis de entradas

	X1 Rel. Compactness	X2 Surface Area	X3 Wall Area	X4 Roof Area
count	768.0	768.0	768.0	768.0
mean	0.764166666667	671.708333333	318.5	176.604166667
std	0.105777475875	88.0861160559	43.626481438	45.1659502229
min	0.62	514.5	245.0	110.25
25%	0.6825	606.375	294.0	140.875
50%	0.75	673.75	318.5	183.75
75%	0.83	741.125	343.0	220.5
max	0.98	808.5	416.5	220.5

	X5 Overall Height	X6 Orientation	X7 Glazing Area	X8 Glazing Area Distr.
count	768.0	768.0	768.0	768.0
mean	5.25	3.5	0.234375	2.8125
std	1.75114043675	1.11876258706	0.133220562915	1.55095966422
min	3.5	2.0	0.0	0.0
25%	3.5	2.75	0.1	1.75
50%	5.25	3.5	0.25	3.0
75%	7.0	4.25	0.4	4.0
max	7.0	5.0	0.4	5.0

Figura 2.2: Estatísticas das variáveis de entradas

Olhando para a valor media das variáveis na figura, nos podemos observar uma diferencia de escala entre as variaveis. Na verdade, a maioria dessas variáveis não têm a mesma unidade:

- $0 < X1 < 1$ sem unidades
- X2, X3, X4 em metros quadrados
- X5 em metros
- X7 em percentagem

- X8 em metros quadrados

Quando nos vamos comparar essas variáveis entre elas, nos vamos ter que padronizar elas. O escolho da métrica de Z-score foi feito, embora que a padronização min-max pudesse ser feita, devido ao fato que a amostra não tem outliers como nos vamos ver na secção seguinte.

Formula de padronização Z-score applicada:

$$\hat{X}_i(t) = \frac{X_i(t) - \bar{X}_i}{\hat{\sigma}_i} \tag{2.1}$$

2.2.2 Variáveis de saídas

	y1 Heating Load	y2 Cooling Load
count	768.0	768.0
mean	22.3071953125	24.5877604167
std	10.0902039702	9.51330556233
min	6.01	10.9
25%	12.9925	15.62
50%	18.95	22.08
75%	31.6675	33.1325
max	43.1	48.03

Figura 2.3: Estatísticas das variáveis de saídas

As variáveis de saída aparecem bastante semelhantes. Nos vamos verificar isso com os histogramas.

2.3 Detecção de outliers

2.3.1 Box-plot

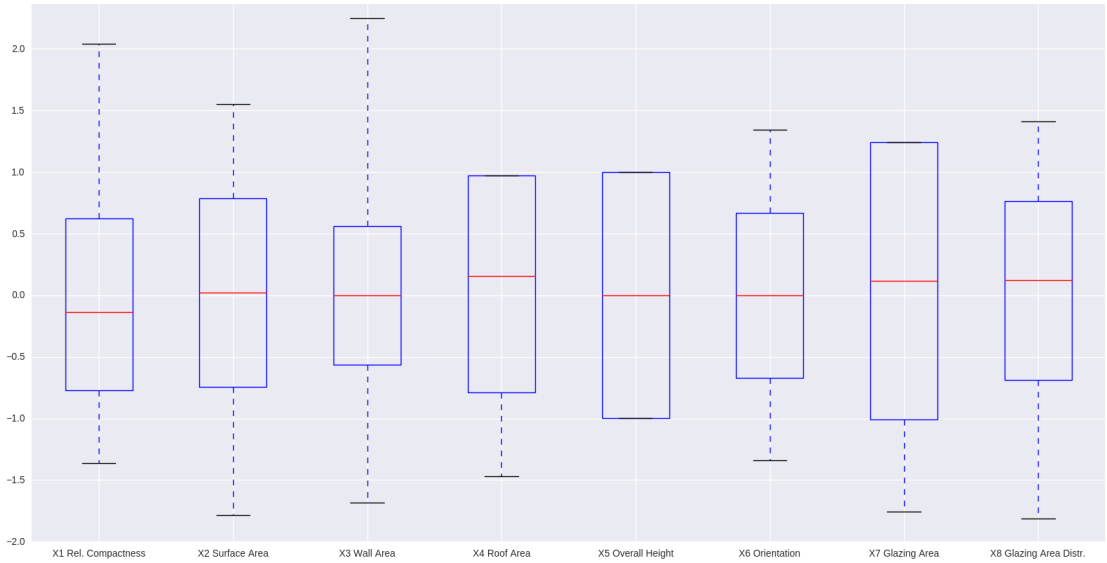

Figura 2.4: Box-plot das variáveis de entradas Z-padronizadas

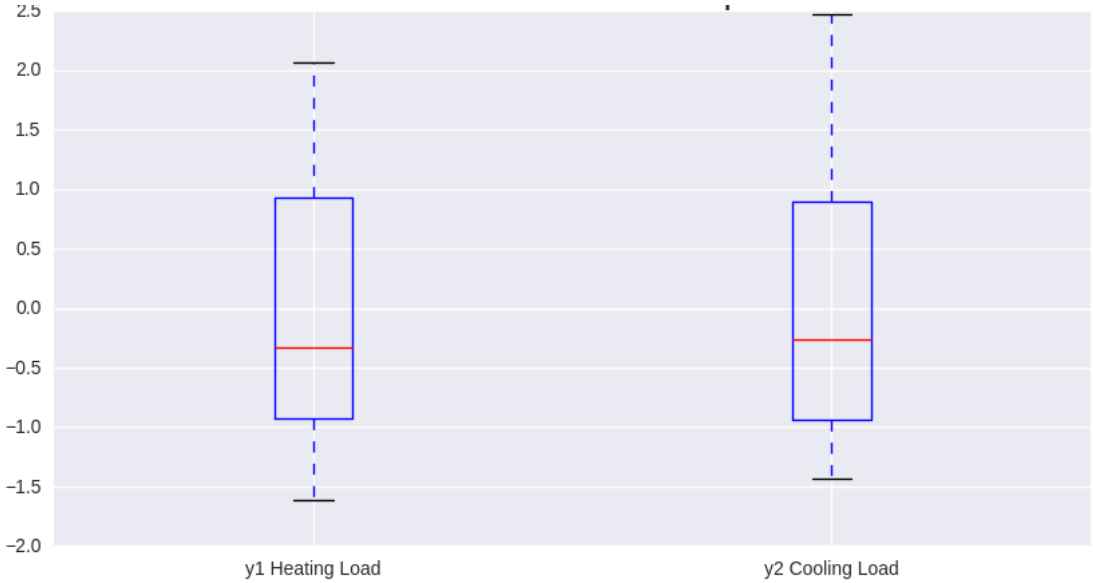

Figura 2.5: Box-plot das variáveis de saídas Z-padronizadas

Pela leitura desses gráficos, nenhuma das variáveis aparecem conter valores aberantes. Mas deve ser considerado todas as variáveis simultaneamente com os métodos baseados em distancia para ser capaz de comprovar esse resultado. Nos vamos nos concentrar apenas nas variáveis de entrada na proxima seção.

2.3.2 A métrica euclideana

A métrica de distancia *Euclidiana* (ou norma L_2) é a extensão da fórmula clássica da geometria para p dimensões:

$$dist_E(v,u) = ||v-u|| = \sum_{i=1}^{p}(v_i - u_i) \tag{2.2}$$

A matriz de distancias obtida é colocada na figura 2.6, e o gráfico das médias de distancias na figura 2.7.

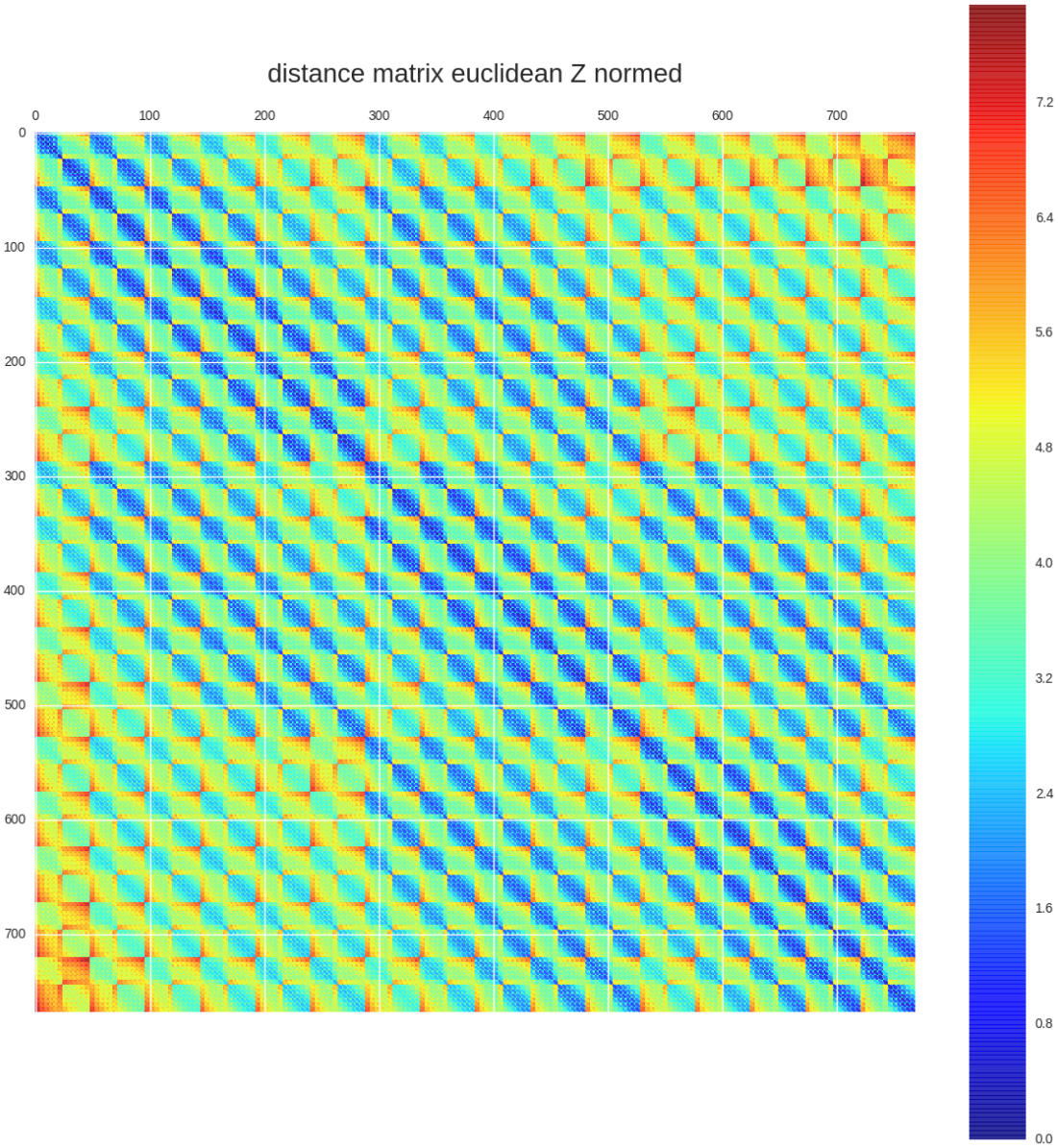

Figura 2.6: Matriz de distancias euclideana com variaveis Z-padronizadas

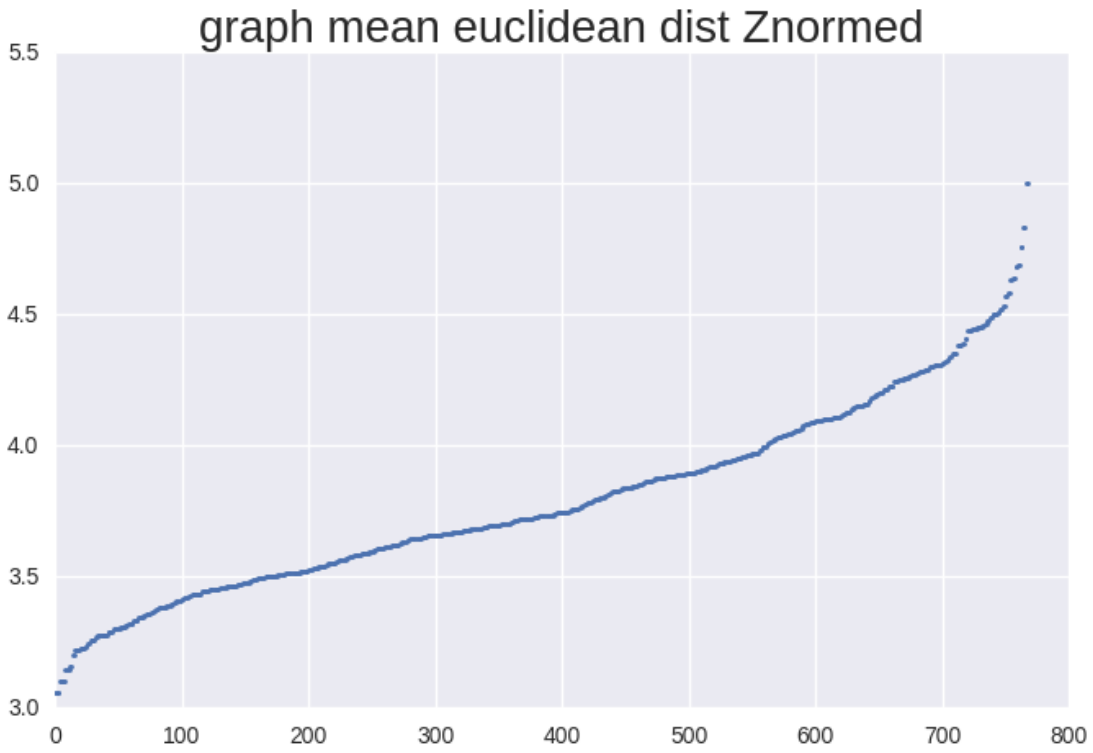

Figura 2.7: Médias de distancias Euclidiana em ordem crescente

2.3.3 A métrica Mahalanobis

Distancia de *Mahalanobis* é a distancia geométrica ponderada pelo inverso da matriz de covariancias estimada no conjunto de dados:

$$dist_{\Sigma}(v,u) = \sqrt{(v-u)\hat{\Sigma}^{-1}(v-u)^{T}} \tag{2.3}$$

A matriz de distancias obtida é colocada na figura 2.8, e o gráfico das médias de distancias na figura 2.9.

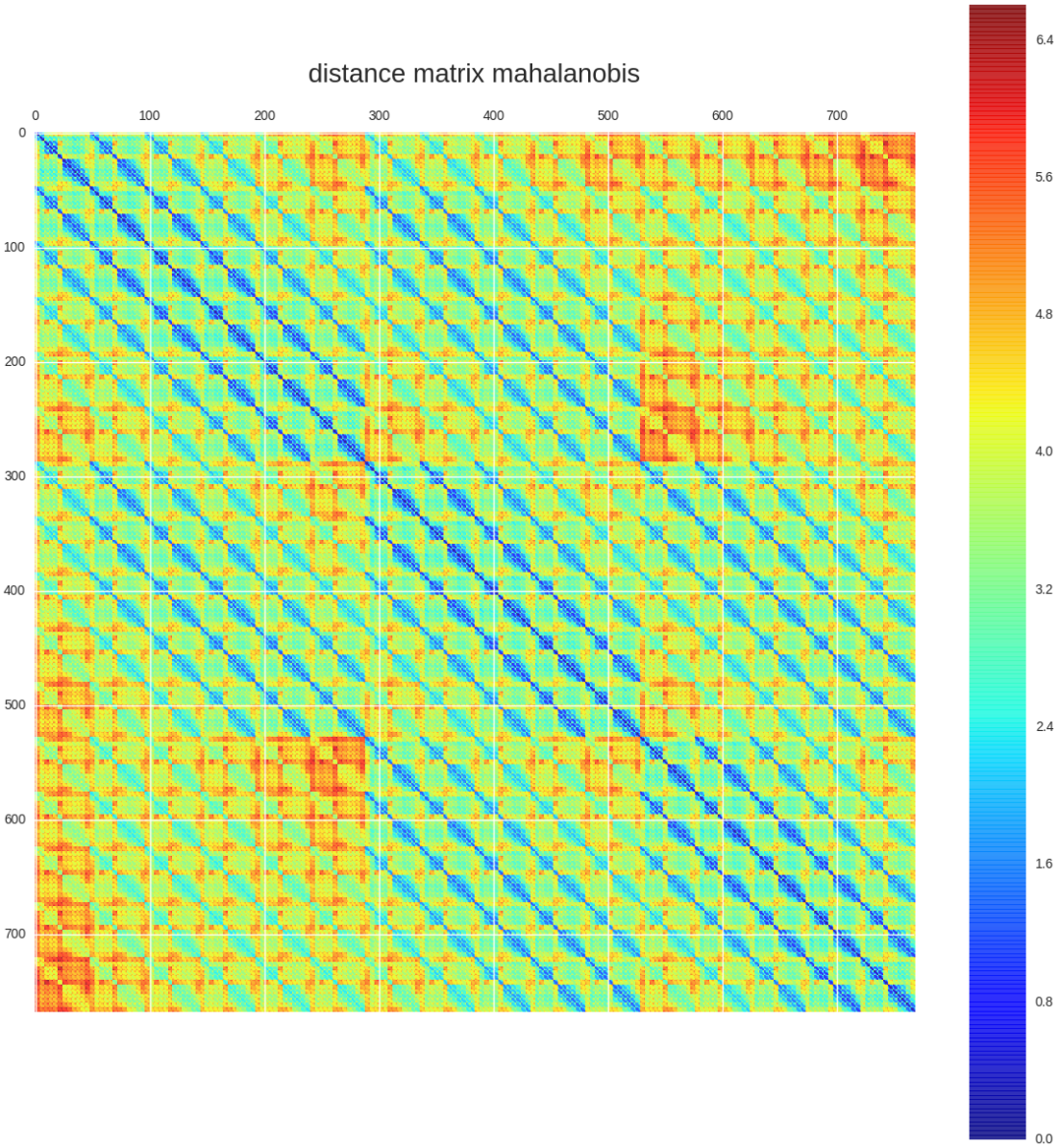

Figura 2.8: Matriz de distancias Mahalanobis

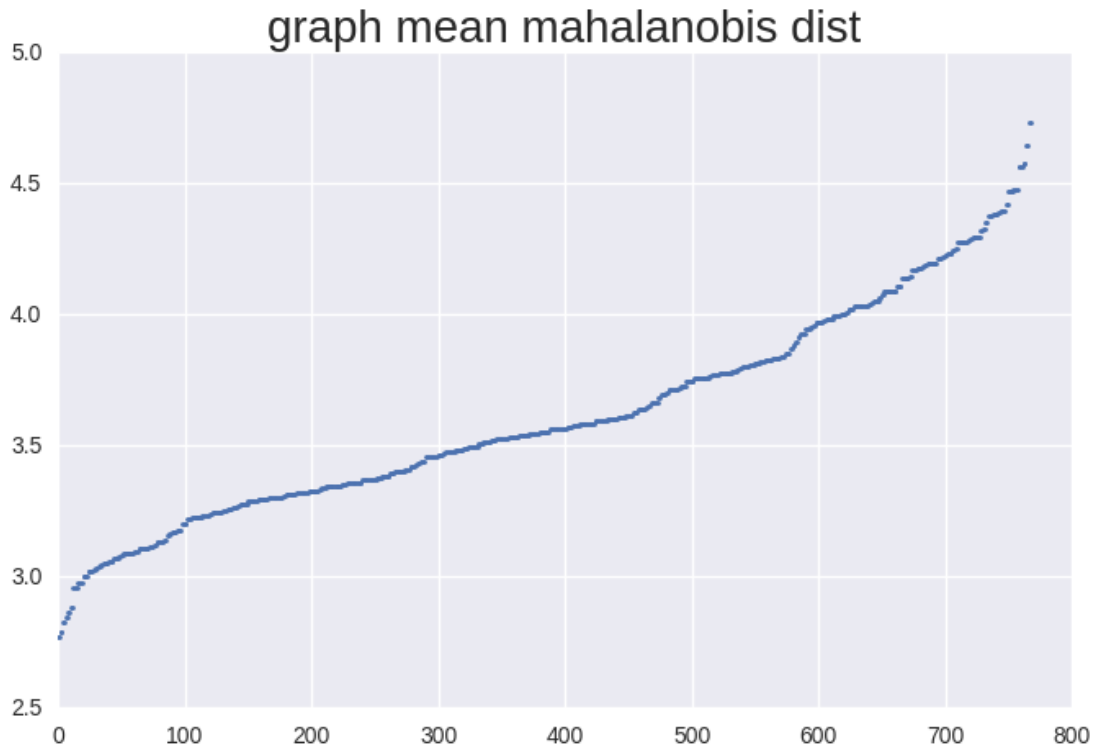

Figura 2.9: Médias de distancias Mahalanobis em ordem crescente

2.3.4 Discussão dos resultados

Como um lembrete, o gráfico das médias ordenadas é obtida pela fórmula:

$$\overline{dist}_i = \frac{1}{N} * \sum_{t=1}^{N} (distance(u_i, v_t)) \tag{2.4}$$

Podemos considerar que não há nenhum registro aberrantes, porque nenhum é fortemente afastado dos outros (seja com a medida euclideana ou a medida Mahalanobis). Além disso, não é surpreendente porque as variáveis de entradas foram selecionadas pelo engenheiro que fiz as simulações, então ele escolheu uma faixa responsável para cada um delas.

É importante de notar que a distancia de Mahalanobis é adequada para medir a separação entre um conjunto de dados com variáveis geradas pela distribuição normal multivariada. No enquanto, nesse conjunto de dado, as distribuições não parecem a ela como nos vamos ver nos histogramas seguintes.

2.4 Distribuições

2.4.1 Histogramas das variáveis não padronizadas

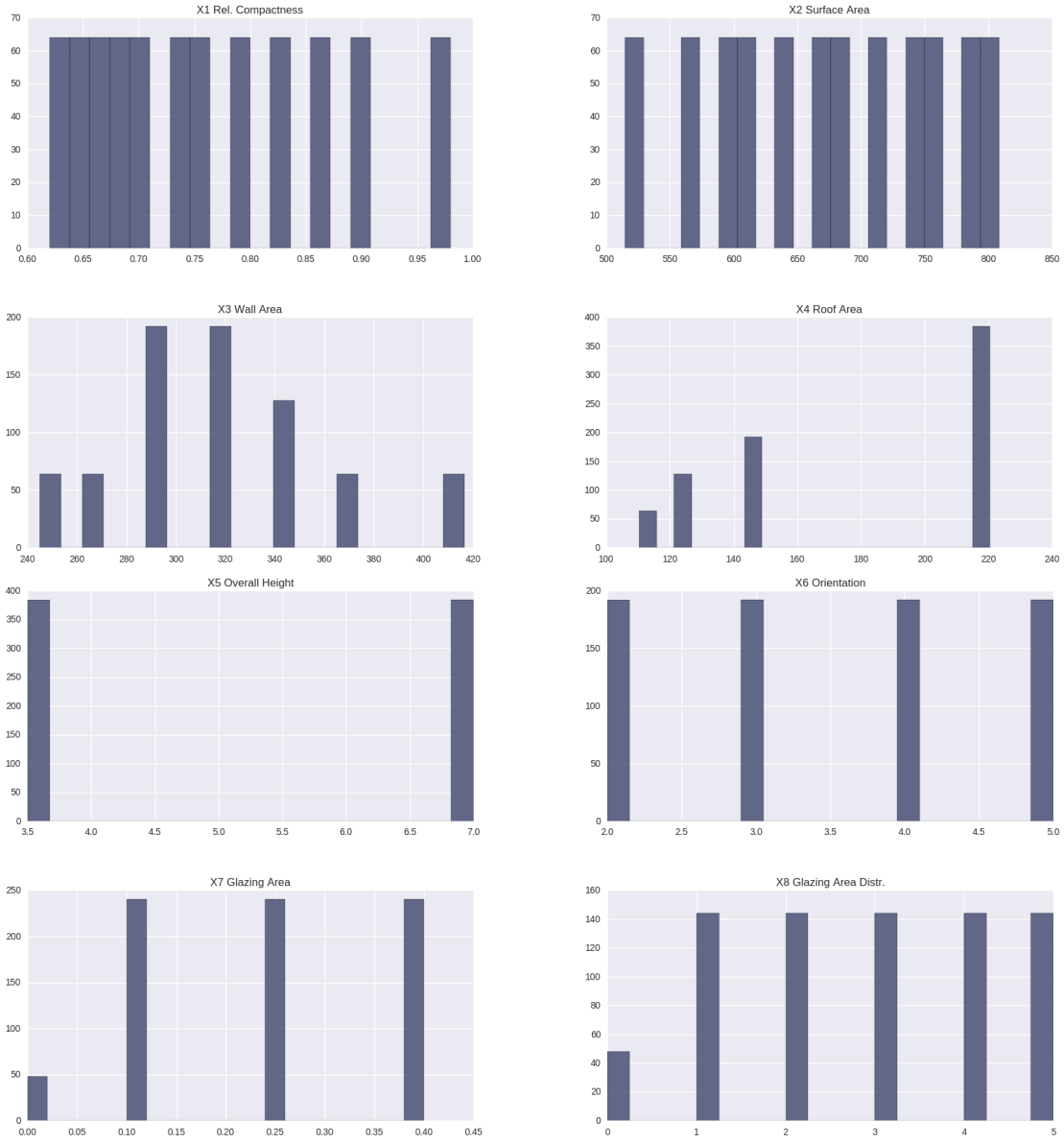

Figura 2.10: Histogramas das variáveis de entradas

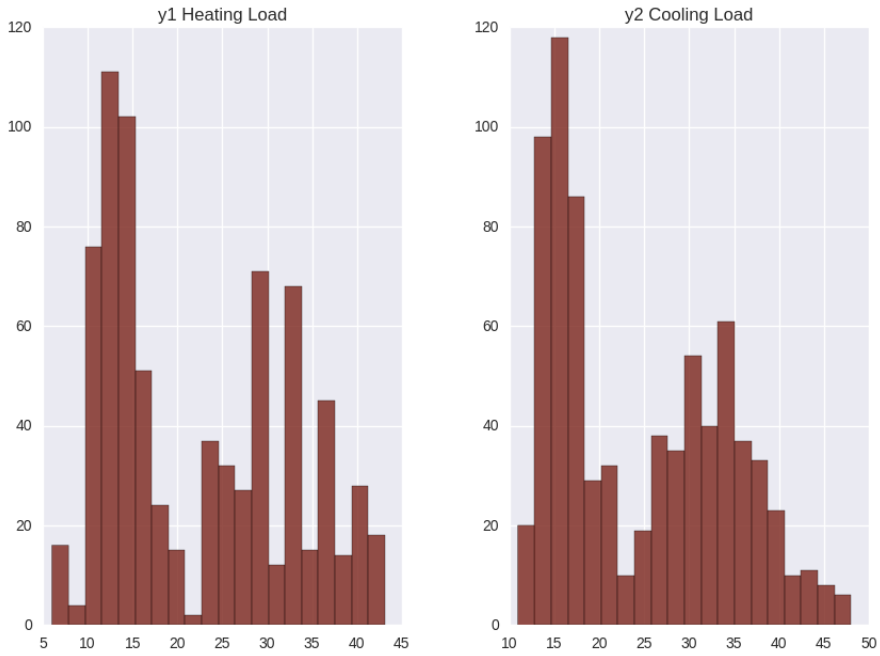

Figura 2.11: Histogramas das variáveis de saídas

Podemos então confirmar que as variáveis de saídas são parecidas. No estudo futuro de regressão, será interessante de somente considerar um delas no primeiro lugar.

Uma outra coisa que deve ser apontada é a forma multimodal das variáveis de saída. Nos podemos já ter em mente que uma regressão linear não vai dar certo.

2.4.2 Gráficos de Projeção

O gráficos de projeção é colocada na figura 2.12

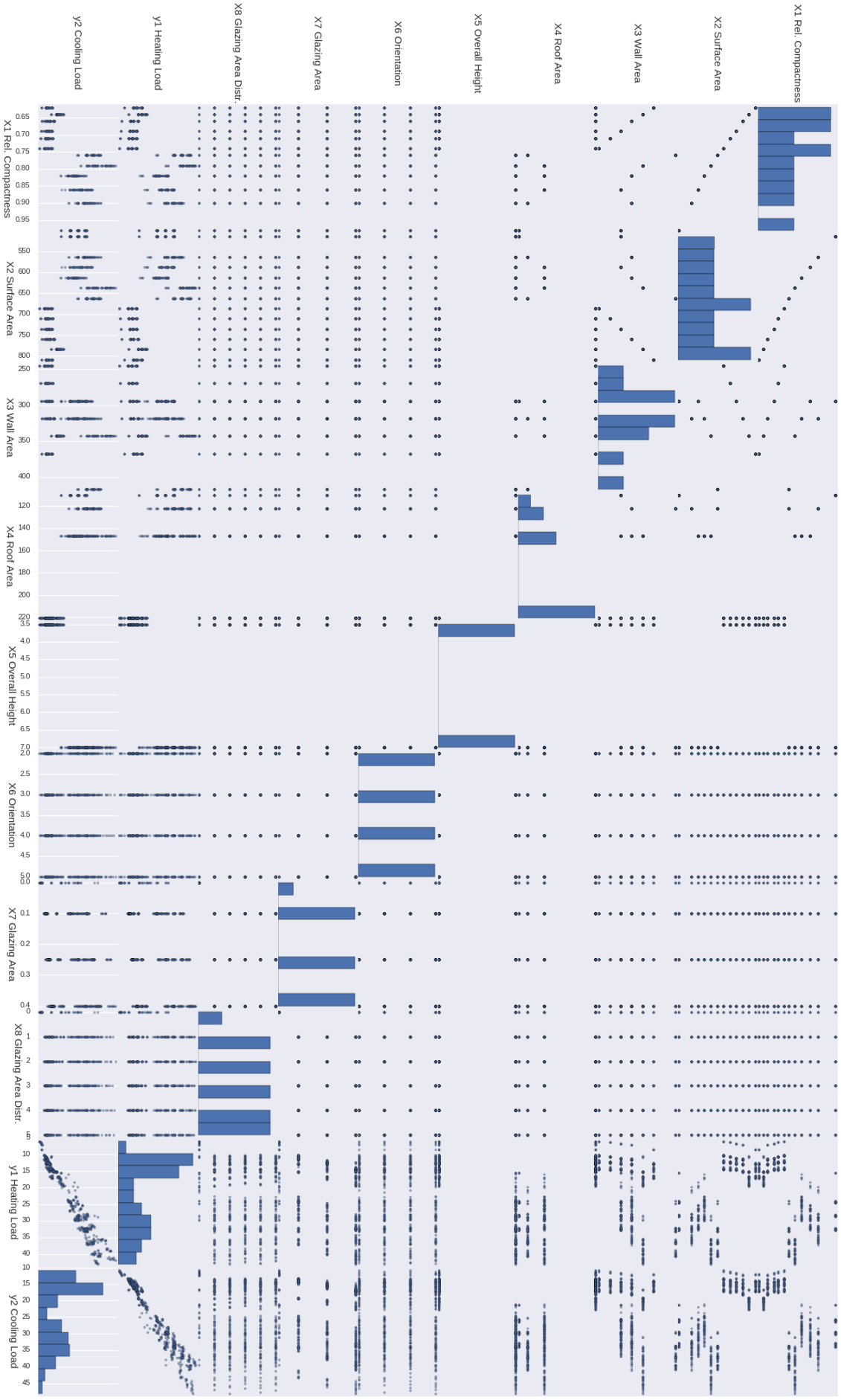

Figura 2.12: Gráficos de Projeção

Parece que as variáveis X1 e X2 são bem correlacionadas, e que a variável X5 não dá muito mais informações de que as outras. Isto é comprovado com a matriz de correlação.

2.5 Matriz de correlação

A matriz de correlação é colocada na figura 2.13

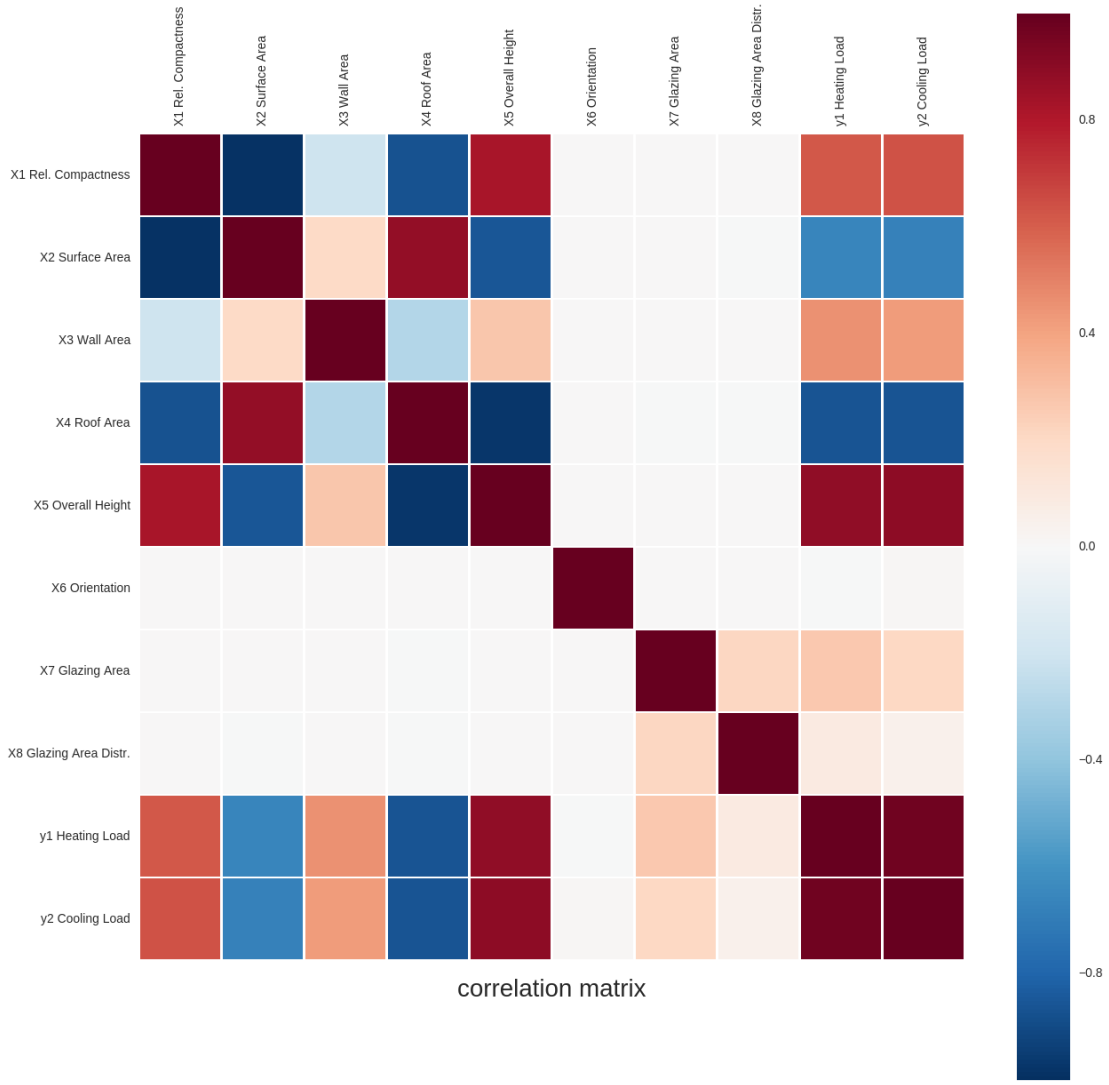

Figura 2.13: Matriz de correlação

Efetivamente, as variáveis X1 (Relative Compactness) e X2 (Surface Area) são inversamente proporcional com um coeficiente de correlação igual a -1. Olhando no papel dos autores, nos podemos encontrar a explicação desse resultado: nos valores escolhidos para as simulações, eles fizeram a hipótese de um volume total dos edifícios constantes. Isto acarreta num relação analítica que liga X1

com X2.

Nos podemos observar também que X4 (Roof Area) é bem correlacionado com X5 (Overall Height), provavelmente devido da mesma hipótese.

Finalmente, as duas variáveis de saída y1 e y2 são fortemente correlacionadas.

3 Discussão

Nos fizemos a analisa de um conjunto de dados com pouco variáveis, que já estava bem condicionada (sem valores ausentes nem outliers). No enquanto, nos vimos que as distribuições não são trívias.

A próxima etapa do projeto vai ser de encontrar um modelo capaz de prever as variáveis y1 (Heat Load) e y2 (Cooling Load) em função das variáveis geométricas dos edifícios. E nos já apontamos que uma regressão linear não parece ser um bom escolho de modelo por causa dessas distribuições foram do comum.

Nos podemos também comentar o fato de que o website UCI colocou esse dataset na categoria de exercícios de classificação. A figura 3.1 mostra a classificação oficial na Europa [5]. O dataset deve ser simplesmente considerado do ponto de vista de classificação energética dos edifícios.

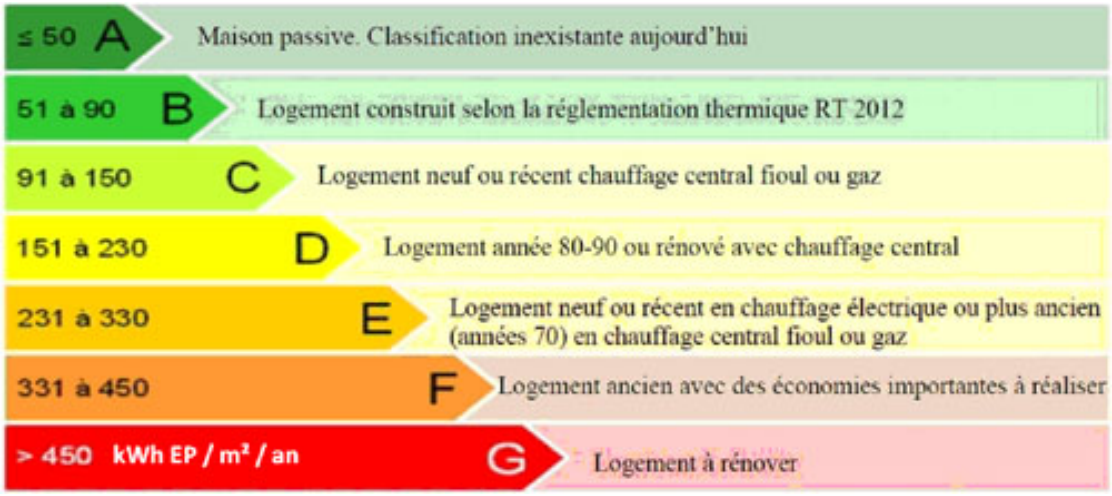

Figura 3.1: Classificação energética dos edifícios na Europa

Lista de Tabelas

Lista de Figuras

Referências

[1] Claude Crampes and Thomas Olivier Léautier. Pour une régulation intelligente de la demande d'électricité. *Les Echos*, 2010.

[2] Journal officiel des Communautés européennes. Directive 2002/91/ce du parlement européen et du conseil sur la performance énergétique des btiments, décembre 2002. http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/.

[3] Tsanas Athanasios and Xifara Angeliki. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, pages Vol. 49, pp. 560–567, 2012.

[4] A. Xifara A. Tsanas. Energy efficiency data set, 2012. https://archive.ics.uci.edu/ml/datasets/Energy+efficiency.

[5] de l'énergie et de la mer Ministère de l'environnement. Diagnostic de performance énergétique, avril 2013. http://www.developpement-durable.gouv.fr/-Diagnostic-de-Performance,855-.html.