Algorithmique et Programmation

Projet : Analyse des séries temporelles
Application au procédé de polissage chimico-mécanique

Explications complémentaires du code, des initiatives et présentation des résultats

EL MAMOUN Kawtar, GACHET Théo
Ecole Nationale Supérieure des Mines de Saint-Etienne

Abstract

L'objectif de ce projet consiste à développer une application en C, qui permet de prétraiter les séries temporelles données afin de prédire et de mieux expliquer l'évolution de la caractéristique cible (i.e. le taux moyen d'élimination du matériel de surface) en utilisant les données historiques.

Contenu du dossier :

```
(dossier) data_trié                 : contient 19 fichiers .csv qui contiennent respectivement les données des 19 séries temporelles triées
(dossier) dataset                   : contient tous les fichiers fournis par l'énoncé (CMP-training-0xx.csv)
(dossier) histogrammes_tracés       : contient les histogrammes calculés à la question 3 pour les 19 séries temporelles
agregation_unique.csv               : contient les WAFER_ID sans répétitions, les occurences sont stockées dans un tableau
agregation.csv                      : contient les WAFER_ID avec répétitions
CMP-training-removalrate.csv        : fichier fourni par l'énoncé pour la question 5
concat.c                            : code permettant de concaténer tous les fichiers texte du dataset en un seul ("resultat.csv")
histogram.py                        : code en Python permettant de visualiser les histogrammes tracés à la question 3 (cf. rapport PDF)
histogrammes.txt                    : affichage console après exécution du code pour la question 3 pour les 19 séries temporelles
histogrammes ksigma.txt             : idem mais après avoir appliqué la règle du ksigma vu à la question 2
main.c                              : code principal
README.txt                          : fichier texte informatif
resultats.csv                       : contient tous les fichiers texte du dataset concaténés
saut_ligne.csv                      : fichier texte utile pour le bon fonctionnement de "concat.c"
statistiques descriptives.txt       : affichage console après exécution du code pour la question 1
statistiques descriptives ksigma.txt : idem mais après avoir appliqué la règle du ksigma vu à la question 2
```

1. Statistiques descriptives centrales et de dispersion : Pour chaque série temporelle, calculer les statistiques descriptives suivantes : moyenne, min, max, écart-type, médiane, ensemble des quartiles. Exporter ces statistiques dans un fichier texte.

```
Statistiques : USAGE_OF_BACKING_FILM
Effectif   : 202084
Moyenne    : 4796.244355
Minimum    : 29.166667
Maximum    : 10299.166667
Ecart_type : 3175.889822
Quartile 1 : 1917.500000
Médiane    : 4865.416667
Quartile 3 : 7480.833333


Statistiques : USAGE_OF_DRESSER
Effectif   : 202084
Moyenne    : 399.445810
Minimum    : 5.185185
Maximum    : 768.888889
Ecart_type : 234.189513
Quartile 1 : 166.666667
Médiane    : 425.925926
Quartile 3 : 604.814815


Statistiques : USAGE_OF_POLISHING_TABLE
Effectif   : 202084
Moyenne    : 171.530049
Minimum    : 0.000000
Maximum    : 357.037037
Ecart_type : 92.576311
Quartile 1 : 90.370370
Médiane    : 170.370370
Quartile 3 : 253.333333
```

```
Statistiques : USAGE_OF_DRESSER_TABLE
Effectif   : 202084
Moyenne    : 2887.778387
Minimum    : 2664.750000
Maximum    : 3205.750000
Ecart_type : 146.771640
Quartile 1 : 2753.500000
Médiane    : 2890.500000
Quartile 3 : 2992.000000


Statistiques : PRESSURIZED_CHAMBER_PRESSURE
Effectif   : 202084
Moyenne    : 53.301020
Minimum    : 0.000000
Maximum    : 188.571429
Ecart_type : 38.756810
Quartile 1 : 0.000000
Médiane    : 73.809524
Quartile 3 : 78.571429


Statistiques : MAIN_OUTER_AIR_BAG_PRESSURE
Effectif   : 202084
Moyenne    : 165.730573
Minimum    : 0.000000
Maximum    : 499.200000
Ecart_type : 131.584549
Quartile 1 : 0.000000
Médiane    : 256.800000
Quartile 3 : 268.800000
```

```
Statistiques : CENTER_AIR_BAG_PRESSURE
Effectif    : 202084
Moyenne     : 42.837811
Minimum     : 0.000000
Maximum     : 138.125000
Ecart_type  : 33.762392
Quartile 1  : 0.000000
Médiane     : 65.937500
Quartile 3  : 72.187500


Statistiques : RETAINER_RING_PRESSURE
Effectif    : 202084
Moyenne     : 1256.773823
Minimum     : 0.000000
Maximum     : 10658.700000
Ecart_type  : 1455.451131
Quartile 1  : 0.000000
Médiane     : 1446.900000
Quartile 3  : 1450.800000


Statistiques : RIPPLE_AIR_BAG_PRESSURE
Effectif    : 202084
Moyenne     : 6.338217
Minimum     : 0.000000
Maximum     : 21.136364
Ecart_type  : 4.963758
Quartile 1  : 0.000000
Médiane     : 9.954545
Quartile 3  : 10.000000


Statistiques : USAGE_OF_MEMBRANE
Effectif    : 202084
Moyenne     : 56.872467
Minimum     : 0.345850
Maximum     : 122.124506
Ecart_type  : 37.658773
Quartile 1  : 22.737154
Médiane     : 57.692688
Quartile 3  : 88.705534


Statistiques : USAGE_OF_PRESSURIZED_SHEET
Effectif    : 202084
Moyenne     : 1438.873435
Minimum     : 8.750000
Maximum     : 3089.750000
Ecart_type  : 952.766961
Quartile 1  : 575.250000
Médiane     : 1459.625000
Quartile 3  : 2244.250000
```

```
Statistiques : SLURRY_FLOW_LINE_A
Effectif    : 202084
Moyenne     : 4.409444
Minimum     : 0.000000
Maximum     : 38.333333
Ecart_type  : 6.640168
Quartile 1  : 2.222222
Médiane     : 2.222222
Quartile 3  : 2.222222


Statistiques : SLURRY_FLOW_LINE_B
Effectif    : 202084
Moyenne     : 0.760229
Minimum     : 0.000000
Maximum     : 12.045455
Ecart_type  : 0.391752
Quartile 1  : 0.909091
Médiane     : 0.909091
Quartile 3  : 0.909091


Statistiques : SLURRY_FLOW_LINE_C
Effectif    : 202084
Moyenne     : 266.341624
Minimum     : 0.000000
Maximum     : 1072.400000
Ecart_type  : 211.940081
Quartile 1  : 0.000000
Médiane     : 422.800000
Quartile 3  : 442.400000


Statistiques : WAFER_ROTATION
Effectif    : 202084
Moyenne     : 4796.198493
Minimum     : 29.166667
Maximum     : 10299.166667
Ecart_type  : 3175.891665
Quartile 1  : 1917.500000
Médiane     : 4865.000000
Quartile 3  : 7480.833333


Statistiques : STAGE_ROTATION
Effectif    : 202084
Moyenne     : 52.929722
Minimum     : 0.000000
Maximum     : 263.552632
Ecart_type  : 92.143500
Quartile 1  : 0.000000
Médiane     : 0.000000
Quartile 3  : 66.052632
```

```
Statistiques : HEAD_ROTATION
Effectif    : 202084
Moyenne     : 160.010134
Minimum     : 118.400000
Maximum     : 192.000000
Ecart_type  : 6.813327
Quartile 1  : 156.800000
Médiane     : 160.000000
Quartile 3  : 160.000000


Statistiques : DRESSING_WATER_STATUS
Effectif    : 202084
Moyenne     : 0.430568
Minimum     : 0.000000
Maximum     : 1.000000
Ecart_type  : 0.495156
Quartile 1  : 0.000000
Médiane     : 0.000000
Quartile 3  : 1.000000
```

```
Statistiques : EDGE_AIR_BAG_PRESSURE
Effectif    : 202084
Moyenne     : 30.538789
Minimum     : 0.000000
Maximum     : 141.515152
Ecart_type  : 24.186933
Quartile 1  : 0.000000
Médiane     : 43.939394
Quartile 3  : 48.484848
```

2. Détection des observations atypiques : Pour chaque série temporelle, détecter et supprimer les observations atypiques en appliquant la règle de kσ (où σ représente l'écart-type), qui consiste à supprimer les observations dont la valeur est supérieure (resp. inférieure) à kσ (resp. −kσ), k = {2, 3}.

```
Statistiques : USAGE_OF_BACKING_FILM
Effectif    : 190638
Moyenne     : 4487.995060
Minimum     : 29.166667
Maximum     : 10299.166667
Ecart_type  : 3001.974340
Quartile 1  : 1767.083333
Médiane     : 4529.166667
Quartile 3  : 7079.166667
```

3. Distributions empiriques : Pour chaque série temporelle, construire l'histogramme associé, i.e. agréger les observations en groupes d'intervalles égaux (bins) et calculer la fréquence des observations dans chaque bin. Fixer un nombre de bins par défaut. (5 ici)

```
--- Génération Histogramme : USAGE_OF_BACKING_FILM ---

Maximum = 10299.166667
Minimum = 29.166667
Pas     = 2054.000000

On obtient les bins suivants :
[29.166667 ; 2083.166667] de fréquence 57755 / 202084
[2083.166667 ; 4137.166667] de fréquence 33868 / 202084
[4137.166667 ; 6191.166667] de fréquence 38884 / 202084
[6191.166667 ; 8245.166667] de fréquence 32606 / 202084
[8245.166667 ; 10299.166667] de fréquence 38971 / 202084
```

```
--- Génération Histogramme : USAGE_OF_DRESSER_TABLE ---

Maximum = 3205.750000
Minimum = 2664.750000
Pas     = 108.200000

On obtient les bins suivants :
[2664.750000 ; 2772.950000] de fréquence 55983 / 202084
[2772.950000 ; 2881.150000] de fréquence 39172 / 202084
[2881.150000 ; 2989.350000] de fréquence 55185 / 202084
[2989.350000 ; 3097.550000] de fréquence 29129 / 202084
[3097.550000 ; 3205.750000] de fréquence 22458 / 202084
[3205.750000 ; 3313.950000] de fréquence 157 / 202084
```

```
--- Génération Histogramme : USAGE_OF_DRESSER ---

Maximum = 768.888889
Minimum = 5.185185
Pas     = 152.740741

On obtient les bins suivants :
[5.185185 ; 157.925926] de fréquence 44203 / 202084
[157.925926 ; 310.666667] de fréquence 41070 / 202084
[310.666667 ; 463.407407] de fréquence 20458 / 202084
[463.407407 ; 616.148148] de fréquence 50933 / 202084
[616.148148 ; 768.888889] de fréquence 45420 / 202084

--- Génération Histogramme : USAGE_OF_POLISHING_TABLE --

Maximum = 357.037037
Minimum = 0.000000
Pas     = 71.407407

On obtient les bins suivants :
[0.000000 ; 71.407407] de fréquence 37788 / 202084
[71.407407 ; 142.814815] de fréquence 45119 / 202084
[142.814815 ; 214.222222] de fréquence 46778 / 202084
[214.222222 ; 285.629630] de fréquence 45302 / 202084
[285.629630 ; 357.037037] de fréquence 27097 / 202084
```

```
--- Génération Histogramme : PRESSURIZED_CHAMBER_PRESSURE

Maximum = 188.571429
Minimum = 0.000000
Pas     = 37.714286

On obtient les bins suivants :
[0.000000 ; 37.714286] de fréquence 68939 / 202084
[37.714286 ; 75.428571] de fréquence 53411 / 202084
[75.428571 ; 113.142857] de fréquence 74719 / 202084
[113.142857 ; 150.857143] de fréquence 4503 / 202084
[150.857143 ; 188.571429] de fréquence 512 / 202084

--- Génération Histogramme : MAIN_OUTER_AIR_BAG_PRESSURE

Maximum = 499.200000
Minimum = 0.000000
Pas     = 99.840000

On obtient les bins suivants :
[0.000000 ; 99.840000] de fréquence 75821 / 202084
[99.840000 ; 199.680000] de fréquence 1893 / 202084
[199.680000 ; 299.520000] de fréquence 120228 / 202084
[299.520000 ; 399.360000] de fréquence 680 / 202084
[399.360000 ; 499.200000] de fréquence 3462 / 202084
```

```
--- Génération Histogramme : CENTER_AIR_BAG_PRESSURE ---

Maximum = 138.125000
Minimum = 0.000000
Pas     = 27.625000

On obtient les bins suivants :
[0.000000 ; 27.625000] de fréquence 75876 / 202084
[27.625000 ; 55.250000] de fréquence 3760 / 202084
[55.250000 ; 82.875000] de fréquence 118063 / 202084
[82.875000 ; 110.500000] de fréquence 3680 / 202084
[110.500000 ; 138.125000] de fréquence 705 / 202084

--- Génération Histogramme : RETAINER_RING_PRESSURE ---

Maximum = 10658.700000
Minimum = 0.000000
Pas     = 2131.740000

On obtient les bins suivants :
[0.000000 ; 2131.740000] de fréquence 185867 / 202084
[2131.740000 ; 4263.480000] de fréquence 6799 / 202084
[4263.480000 ; 6395.220000] de fréquence 1355 / 202084
[6395.220000 ; 8526.960000] de fréquence 7762 / 202084
[8526.960000 ; 10658.700000] de fréquence 301 / 202084

--- Génération Histogramme : RIPPLE_AIR_BAG_PRESSURE ---

Maximum = 21.136364
Minimum = 0.000000
Pas     = 4.227273

On obtient les bins suivants :
[0.000000 ; 4.227273] de fréquence 75486 / 202084
[4.227273 ; 8.454545] de fréquence 2667 / 202084
[8.454545 ; 12.681818] de fréquence 119696 / 202084
[12.681818 ; 16.909091] de fréquence 779 / 202084
[16.909091 ; 21.136364] de fréquence 3456 / 202084
```

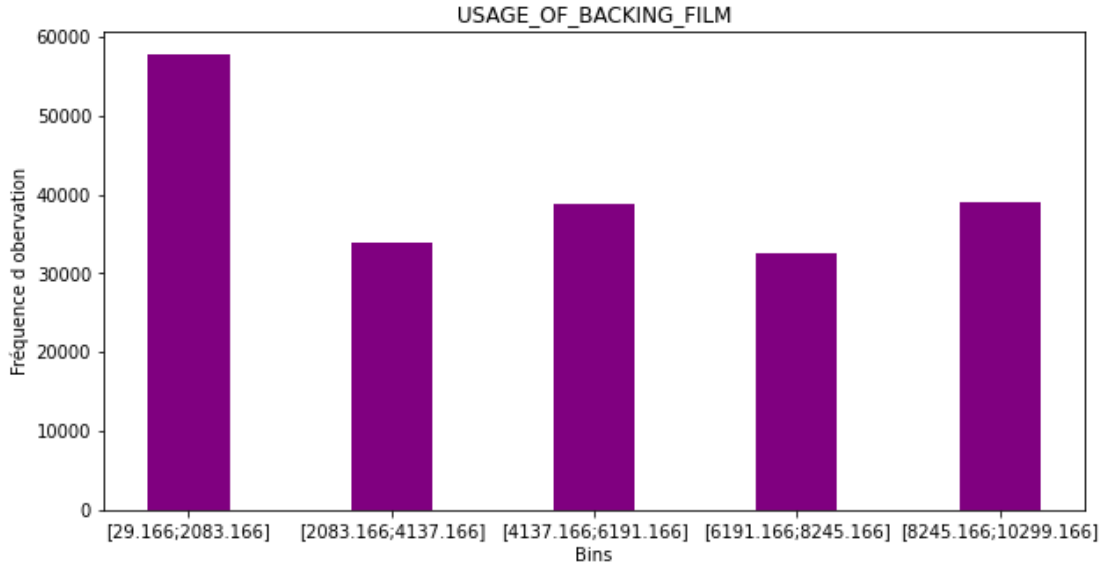
```
--- Génération Histogramme : USAGE_OF_MEMBRANE ---

Maximum = 122.124506
Minimum = 0.345850
Pas     = 24.355731

On obtient les bins suivants :
[0.345850 ; 24.701581] de fréquence 57755 / 202084
[24.701581 ; 49.057312] de fréquence 33868 / 202084
[49.057312 ; 73.413043] de fréquence 38884 / 202084
[73.413043 ; 97.768775] de fréquence 32606 / 202084
[97.768775 ; 122.124506] de fréquence 38971 / 202084

--- Génération Histogramme : USAGE_OF_PRESSURIZED_SHEET --

Maximum = 3089.750000
Minimum = 8.750000
Pas     = 616.200000

On obtient les bins suivants :
[8.750000 ; 624.950000] de fréquence 57755 / 202084
[624.950000 ; 1241.150000] de fréquence 33868 / 202084
[1241.150000 ; 1857.350000] de fréquence 38884 / 202084
[1857.350000 ; 2473.550000] de fréquence 32606 / 202084
[2473.550000 ; 3089.750000] de fréquence 38971 / 202084

--- Génération Histogramme : SLURRY_FLOW_LINE_A ---

Maximum = 38.333333
Minimum = 0.000000
Pas     = 7.666667

On obtient les bins suivants :
[0.000000 ; 7.666667] de fréquence 175006 / 202084
[7.666667 ; 15.333333] de fréquence 6298 / 202084
[15.333333 ; 23.000000] de fréquence 9318 / 202084
[23.000000 ; 30.666667] de fréquence 11358 / 202084
[30.666667 ; 38.333333] de fréquence 104 / 202084
```

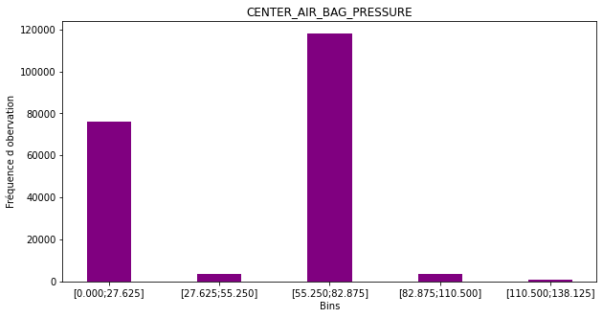
```
--- Génération Histogramme : SLURRY_FLOW_LINE_B ---

Maximum = 12.045455
Minimum = 0.000000
Pas     = 2.409091

On obtient les bins suivants :
[0.000000 ; 2.409091] de fréquence 201957 / 202084
[2.409091 ; 4.818182] de fréquence 34 / 202084
[4.818182 ; 7.227273] de fréquence 20 / 202084
[7.227273 ; 9.636364] de fréquence 22 / 202084
[9.636364 ; 12.045455] de fréquence 51 / 202084

--- Génération Histogramme : SLURRY_FLOW_LINE_C ---

Maximum = 1072.400000
Minimum = 0.000000
Pas     = 214.480000

On obtient les bins suivants :
[0.000000 ; 214.480000] de fréquence 80147 / 202084
[214.480000 ; 428.960000] de fréquence 28232 / 202084
[428.960000 ; 643.440000] de fréquence 93352 / 202084
[643.440000 ; 857.920000] de fréquence 225 / 202084
[857.920000 ; 1072.400000] de fréquence 128 / 202084

--- Génération Histogramme : WAFER_ROTATION ---

Maximum = 10299.166667
Minimum = 29.166667
Pas     = 2054.000000

On obtient les bins suivants :
[29.166667 ; 2083.166667] de fréquence 57756 / 202084
[2083.166667 ; 4137.166667] de fréquence 33868 / 202084
[4137.166667 ; 6191.166667] de fréquence 38884 / 202084
[6191.166667 ; 8245.166667] de fréquence 32606 / 202084
[8245.166667 ; 10299.166667] de fréquence 38970 / 202084
```

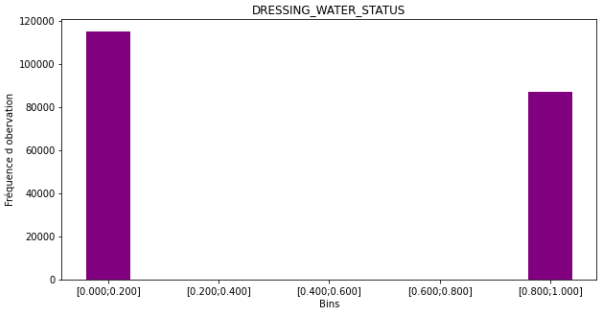
```
--- Génération Histogramme : STAGE_ROTATION ---

Maximum = 263.552632
Minimum = 0.000000
Pas     = 52.710526

On obtient les bins suivants :
[0.000000 ; 52.710526] de fréquence 144016 / 202084
[52.710526 ; 105.421053] de fréquence 10902 / 202084
[105.421053 ; 158.131579] de fréquence 12320 / 202084
[158.131579 ; 210.842105] de fréquence 12363 / 202084
[210.842105 ; 263.552632] de fréquence 22483 / 202084

--- Génération Histogramme : HEAD_ROTATION ---

Maximum = 192.000000
Minimum = 118.400000
Pas     = 14.720000

On obtient les bins suivants :
[118.400000 ; 133.120000] de fréquence 7 / 202084
[133.120000 ; 147.840000] de fréquence 0 / 202084
[147.840000 ; 162.560000] de fréquence 188153 / 202084
[162.560000 ; 177.280000] de fréquence 4568 / 202084
[177.280000 ; 192.000000] de fréquence 9356 / 202084

--- Génération Histogramme : DRESSING_WATER_STATUS ---

Maximum = 1.000000
Minimum = 0.000000
Pas     = 0.200000

On obtient les bins suivants :
[0.000000 ; 0.200000] de fréquence 115073 / 202084
[0.200000 ; 0.400000] de fréquence 0 / 202084
[0.400000 ; 0.600000] de fréquence 0 / 202084
[0.600000 ; 0.800000] de fréquence 0 / 202084
[0.800000 ; 1.000000] de fréquence 87011 / 202084

--- Génération Histogramme : EDGE_AIR_BAG_PRESSURE ---

Maximum = 141.515152
Minimum = 0.000000
Pas     = 28.303030

On obtient les bins suivants :
[0.000000 ; 28.303030] de fréquence 76325 / 202084
[28.303030 ; 56.606061] de fréquence 108222 / 202084
[56.606061 ; 84.909091] de fréquence 16831 / 202084
[84.909091 ; 113.212121] de fréquence 280 / 202084
[113.212121 ; 141.515152] de fréquence 426 / 202084
```

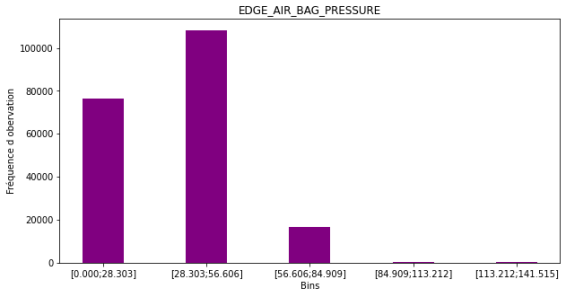
Nous avons également rédigé un script Python afin de pouvoir visualiser nos résultats :

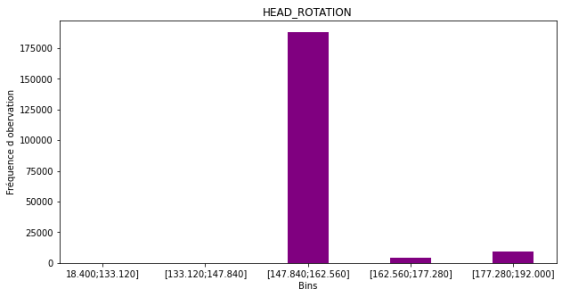
```python
import numpy as np
import matplotlib.pyplot as plt

# programme d'affichage des histogrammes
# des différentes séries temporelles
# dans cet exemple démonstratif,
# nous visualisons USAGE_OF_BACKING_FILM

data = {"[29.166;2083.166]":57755,
        "[2083.166;4137.166]":33868,
        "[4137.166;6191.166]":38884,
        "[6191.166;8245.166]":32606,
        "[8245.166;10299.166]":38971}
donnees = list(data.keys())
valeurs = list(data.values())
fig = plt.figure(figsize = (10, 5))
plt.bar(donnees, valeurs, color ='purple',width = 0.4)
plt.xlabel("Bins")
plt.ylabel("Fréquence d obervation")
plt.title("USAGE_OF_BACKING_FILM")
plt.show()
```

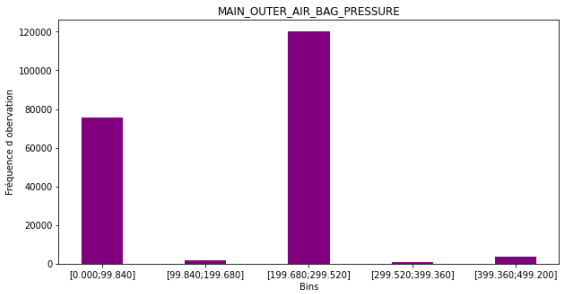

USAGE_OF_BACKING_FILM
Fréquence d obervation
60000
50000
40000
30000
20000
10000
0
[29.166;2083.166]
[2083.166;4137.166]
[4137.166;6191.166]
[6191.166;8245.166]
[8245.166;10299.166]
Bins

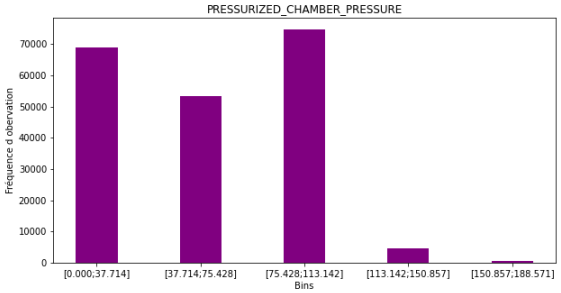

CENTER_AIR_BAG_PRESSURE
Fréquence d obervation
120000
100000
80000
60000
40000
20000
0
[0.000;27.625]
[27.625;55.250]
[55.250;82.875]
[82.875;110.500]
[110.500;138.125]
Bins

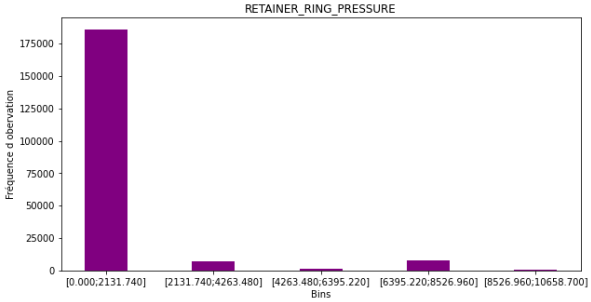

DRESSING_WATER_STATUS
Fréquence d obervation
120000
100000
80000
60000
40000
20000
0
[0.000;0.200]
[0.200;0.400]
[0.400;0.600]
[0.600;0.800]
[0.800;1.000]
Bins

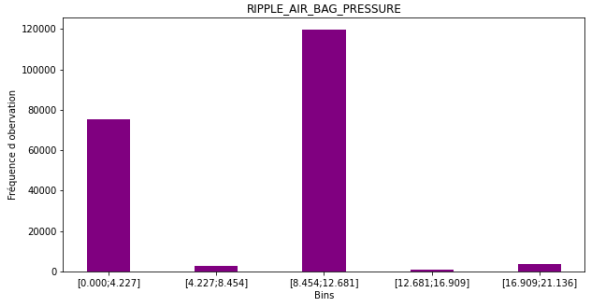

EDGE_AIR_BAG_PRESSURE
Fréquence d obervation
100000
80000
60000
40000
20000
0
[0.000;28.303]
[28.303;56.606]
[56.606;84.909]
[84.909;113.212]
[113.212;141.515]
Bins

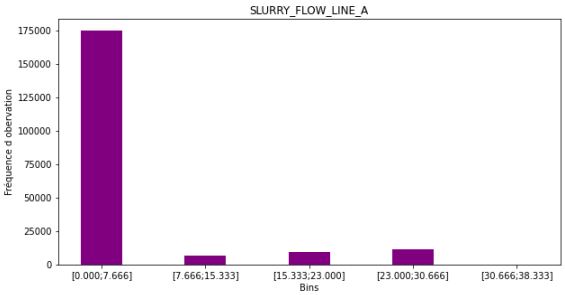

HEAD_ROTATION
Fréquence d obervation
175000
150000
125000
100000
75000
50000
25000
0
18.400;133.120]
[133.120;147.840]
[147.840;162.560]
[162.560;177.280]
[177.280;192.000]
Bins

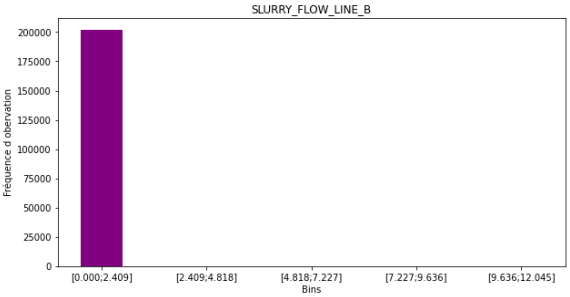

MAIN_OUTER_AIR_BAG_PRESSURE
Fréquence d obervation
120000
100000
80000
60000
40000
20000
0
[0.000;99.840]
[99.840;199.680]
[199.680;299.520]
[299.520;399.360]
[399.360;499.200]
Bins

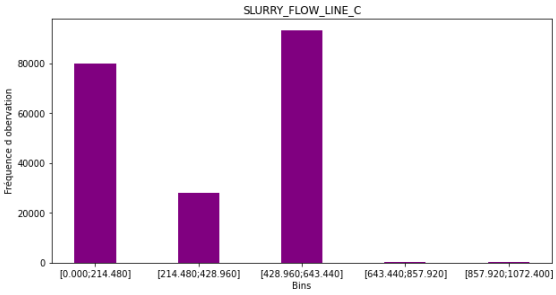

PRESSURIZED_CHAMBER_PRESSURE
Fréquence d obervation
70000
60000
50000
40000
30000
20000
10000
0
[0.000;37.714]
[37.714;75.428]
[75.428;113.142]
[113.142;150.857]
[150.857;188.571]
Bins

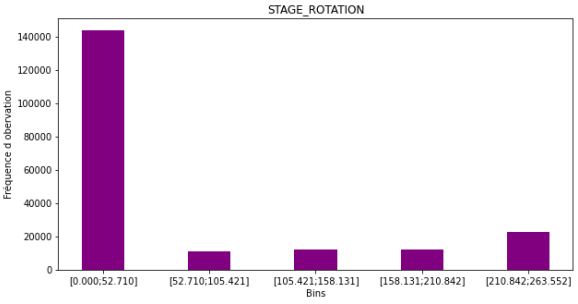

RETAINER_RING_PRESSURE
Fréquence d obervation
175000
150000
125000
100000
75000
50000
25000
0
[0.000;2131.740]
[2131.740;4263.480]
[4263.480;6395.220]
[6395.220;8526.960]
[8526.960;10658.700]
Bins

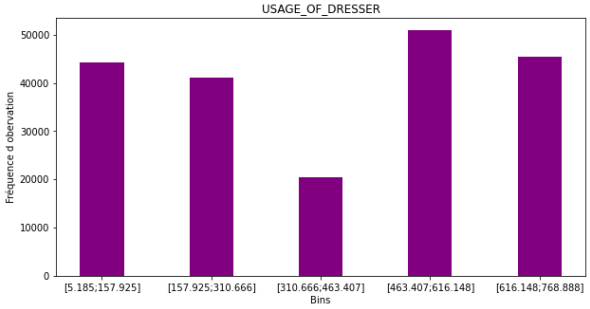

RIPPLE_AIR_BAG_PRESSURE
Fréquence d obervation
120000
100000
80000
60000
40000
20000
0
[0.000;4.227]
[4.227;8.454]
[8.454;12.681]
[12.681;16.909]
[16.909;21.136]
Bins

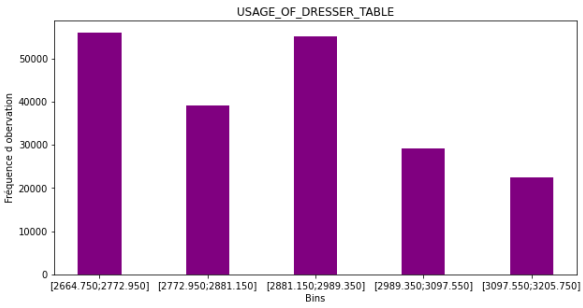

SLURRY_FLOW_LINE_A
Fréquence d obervation
Bins
[0.000;7.666]
[7.666;15.333]
[15.333;23.000]
[23.000;30.666]
[30.666;38.333]

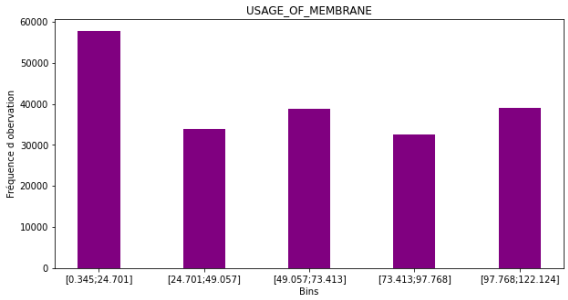

SLURRY_FLOW_LINE_B
Fréquence d obervation
Bins
[0.000;2.409]
[2.409;4.818]
[4.818;7.227]
[7.227;9.636]
[9.636;12.045]

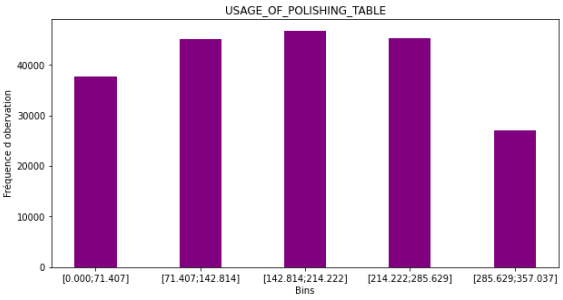

SLURRY_FLOW_LINE_C
Fréquence d obervation
Bins
[0.000;214.480]
[214.480;428.960]
[428.960;643.440]
[643.440;857.920]
[857.920;1072.400]

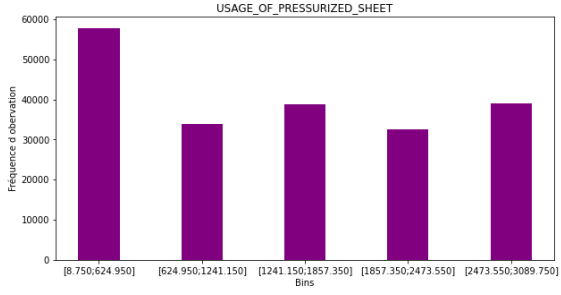

STAGE_ROTATION
Fréquence d obervation
Bins
[0.000;52.710]
[52.710;105.421]
[105.421;158.131]
[158.131;210.842]
[210.842;263.552]

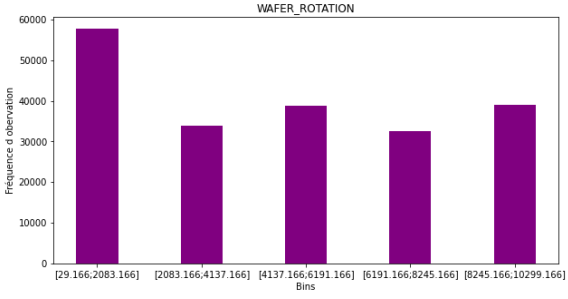

USAGE_OF_DRESSER
Fréquence d obervation
Bins
[5.185;157.925]
[157.925;310.666]
[310.666;463.407]
[463.407;616.148]
[616.148;768.888]


USAGE_OF_DRESSER_TABLE
Fréquence d obervation
Bins
[2664.750;2772.950]
[2772.950;2881.150]
[2881.150;2989.350]
[2989.350;3097.550]
[3097.550;3205.750]


USAGE_OF_MEMBRANE
Fréquence d obervation
Bins
[0.345;24.701]
[24.701;49.057]
[49.057;73.413]
[73.413;97.768]
[97.768;122.124]


USAGE_OF_POLISHING_TABLE
Fréquence d obervation
Bins
[0.000;71.407]
[71.407;142.814]
[142.814;214.222]
[214.222;285.629]
[285.629;357.037]


USAGE_OF_PRESSURIZED_SHEET
Fréquence d obervation
Bins
[8.750;624.950]
[624.950;1241.150]
[1241.150;1857.350]
[1857.350;2473.550]
[2473.550;3089.750]


WAFER_ROTATION
Fréquence d obervation
Bins
[29.166;2083.166]
[2083.166;4137.166]
[4137.166;6191.166]
[6191.166;8245.166]
[8245.166;10299.166]

Lorsque nous appliquons la règle de kσ (question 2), les résultats changent légèrement :

```
--- Génération histogramme : USAGE_OF_BACKING_FILM ---

Maximum = 10299.166667
Minimum = 29.166667
Pas     = 2054.000000

On obtient les bins suivants :
[29.166667;2083.166667] de fréquence 57755 / 202084
[2083.166667;4137.166667] de fréquence 33868 / 202084
[4137.166667;6191.166667] de fréquence 38884 / 202084
[6191.166667;8245.166667] de fréquence 32606 / 202084
[8245.166667;10299.166667] de fréquence 27524 / 202084

--- Génération histogramme : USAGE_OF_DRESSER ---

Maximum = 768.888889
Minimum = 5.185185
Pas     = 152.740741

On obtient les bins suivants :
[5.185185;157.925926] de fréquence 44203 / 202084
[157.925926;310.666667] de fréquence 41070 / 202084
[310.666667;463.407407] de fréquence 20458 / 202084
[463.407407;616.148148] de fréquence 50933 / 202084
[616.148148;768.888889] de fréquence 36976 / 202084
```

```
--- Génération histogramme : USAGE_OF_POLISHING_TABLE -

Maximum = 357.037037
Minimum = 0.000000
Pas     = 71.407407

On obtient les bins suivants :
[0.000000;71.407407] de fréquence 37785 / 202084
[71.407407;142.814815] de fréquence 45118 / 202084
[142.814815;214.222222] de fréquence 46775 / 202084
[214.222222;285.629630] de fréquence 42461 / 202084
[285.629630;357.037037] de fréquence 13558 / 202084

--- Génération histogramme : USAGE_OF_DRESSER_TABLE ---

Maximum = 3205.750000
Minimum = 2664.750000
Pas     = 108.200000

On obtient les bins suivants :
[2664.750000;2772.950000] de fréquence 27991 / 202084
[2772.950000;2881.150000] de fréquence 19586 / 202084
[2881.150000;2989.350000] de fréquence 27593 / 202084
[2989.350000;3097.550000] de fréquence 14564 / 202084
[3097.550000;3205.750000] de fréquence 11229 / 202084
[3205.750000;3313.950000] de fréquence 79 / 202084
```

4. Extraction des caractéristiques agrégées par wafer : Agréger les séries temporelles par wafer. Utiliser comme critère d'agrégation la moyenne. Sauvegarder les données agrégées dans un fichier texte.

Pour cette question, nous avons d'abord concaténé les 58 fichiers du dataset grâce à notre fonction concat() que vous pouvez trouver dans le dossier joint à ce rapport. Ensuite, nous avons regroupé tous les WAFER_ID comme l'énoncé le demande. Puisque le critère de corrélation imposé est la moyenne, nous allons calculer, pour chaque wafer, les moyennes de chaque série temporelle.

Ainsi, notre fonction regrouper_wafer1() créé un fichier contenant les WAFER_ID pour retrouver les redondances et les répétitions. Ensuite, la sortie de cette fonction est donnée en entrée de regrouper_wafer2() qui renvoie un fichier contenant des WAFER_ID uniques ainsi qu'un tableau contenant le nombre d'occurrences de chaque ID dans le fichier concaténé. Ensuite, notre fonction tronquer() tronque le fichier agregation_unique.csv afin d'effectuer les moyennes sur les wafers.

NB : Le code a été intégralement simplifié et commenté, ainsi que les documentations. Nous vous conseillons néanmoins de l'ouvrir avec des éditeurs de code tels que Visual Studio Code car certaines fonctions sont, de par les contraintes du sujet, de « lourds » blocs de variables. Ainsi, certains éditeurs de texte permettent de réduire les boucles et les fonctions (cf. image ci-dessous).

```
585 > int stock_USAGE_OF_DRESSER(char *training, Liste *liste)···
651 > int stock_USAGE_OF_POLISHING_TABLE(char *training, Liste *liste)···
717 > int stock_USAGE_OF_DRESSER_TABLE(char *training, Liste *liste)···
```

Les commentaires ont été disposés sur le code de manière à le rendre le plus clair possible et la fonction main(), située à la toute fin du fichier main.c, contient les instructions permettant de répondre aux questions 1 à 4 en retirant les commentaires.

Les questions 5 et 6 n'ont pu être abordées que partiellement du fait d'un manque de temps, elles seront néanmoins approfondies de notre côté afin de pleinement saisir l'enjeu du sujet sur le traitement des données de grande dimension et la prédiction par régression linéaire.