POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

MTH8302

Modèles de régression et d'analyse de la variance

Devoir 4

distribution : samedi 09 mars 2024

remise (au plus tard) : lundi 30 avril 2024 - 12h00

Ce travail est réalisé individuellement par chaque étudiant inscrit au cours.
Chaque étudiant le fait SEUL sans demander de l'aide à d'autres.
En apposant sa signature ci-dessous, l'étudiant (e) certifie sur son honneur avoir effectué ce travail SEUL.
L'obtention des résultats présentés et la rédaction de ce travail ne fait l'objet d'aucun plagiat, partiel ou total.

Information concernant le plagiat à Polytechnique : http://www.polymtl.ca/etudes/ppp/index.php

Exigences pour la rédaction du rapport : consulter la page 2 du plan de cours.

https://cours.polymtl.ca/mth6301/mth8302-Cours&Plus/2024-MTH8302-ch00-PlanCours-hiver2024.pdf

Compléter l'information suivante et transmettez cette page comme la page 1 de votre rapport de devoir.
Une copie de cette page est disponible sur le site du cours.

MTH8302 Modèles de régression et d'analyse de variance

NOM _____ **PRÉNOM** _____

MATRICULE _____ **SIGNATURE** _____

Transmettez votre rapport par courriel à bernard.clement@polymtl.ca

Nom obligatoire du fichier a transmettre : FFFF_mmm_2024_MTH8302_devoir4.pdf

FFFF = nom de famille **mmm = matricule**

Le devoir4 compte pour 35% pour l'évaluation de la note finale du cours.

TABLEAU Devoir4	valeur	thème
MTH8302-Exer-34-partie A Rapport lecture sur le Contrôle Statistique de Procédé (SPC)	20	Examen des données avec SPC Identification et séparation des données
MTH8302-Exer 34-partie B Étude d'un cas avec et sans SPC	60	Modele MARS avec et sans SPC Modèle SANN avec et sans SPC
Qualité générale	20	
TOTAL	100	

Tous les fichiers pour la réalisation du devoir4 sont disponibles sur le site WEB du cours :

https://cours.polymtl.ca/mth6301/MTH8302.htm

Les liens pour les fichiers de lecture sont donnés à la page2.

Les fichiers de données sont donnés en format STATISTICA (sta) et en format EXCEL (xlsx)

https://cours.polymtl.ca/mth6301/mth8302-Cours&Plus/BeverageManufacturing%20(37vX2361c).sta

https://cours.polymtl.ca/mth6301/mth8302-Cours&Plus/BeverageManufacturing%20(37vX2361c).xlsx

Avant propos

J'ai annoncé que le devoir 4 aurait la forme d'une étude de cas. Parallèlement, j'ai décidé d'enrichir vos connaissances en introduisant une méthode que je soupçonne vous n'avez jamais vu dans vos formations antérieures en ingénierie et en méthodes statistiques. La méthode en question porte le nom de contrôle statistique de procédé. Elle est généralement représentée par le sigle SPC, l'abréviation de Statistical Process Control. J'ai pensé combiner la méthode du SPC avec la méthode de développement de modèles de prédiction sur un ensemble de données industrielles. Le projet que je vous soumets est unique au monde et original dans le domaine de l'apprentissage statistique (Machine Learning). A ma connaissance, le concept du SPC est totalement absent dans la littérature du Machine Learning. Dans la première partie (Exer-34), il faudra prendre connaissance de deux documents.

Document1 (4 pages) : résumé d'un article de Donald J. Wheeler
https://cours.polymtl.ca/mth6301/mth8302-Cours&Plus/Clement/Clement-Etudes_observationnelles_Etudes_experimentales.pdf

Document2 (31 pages) : présentation du contrôle statistique des processus (SPC)
https://cours.polymtl.ca/mth6301/mth8302-Cours&Plus/Clement-ControleStatistiqueProcessus.pdf

La méthodologie du SPC et sa mise en œuvre avec Statistica est proposée dans le document2. Des données réelles seront employées dans les deux parties (A et B) du devoir. Il s'agit d'une extension au cours MTH8302 et elle n'en fait pas partie dans sa description officielle. De même, les réseaux de neurones et autres méthodes du Data Mining (Machine Learning).

Le SPC ou le contrôle statistique des procédés pourrait aussi s'appeler, étude du comportement des procédés. Le SPC est un état d'esprit plutôt qu'une machine à calculs statistiques. La méthode permet d'apprendre et de comprendre le comportement d'un système ou processus. Le SPC vise à augmenter le niveau de connaissances et d'établir des relations causales. Ceci sans développer une artillerie de calculs mathématiques associés aux méthodes avancées de modélisation statistique. La stabilité statistique d'un processus est en liaison directe avec l'homogénéité des données. Cette question est d'une certaine manière reliée à la présence possible et l'identification de données aberrantes. Le SPC vise à déterminer les causes spéciales de la présence de ces données associées. Ultimement on veut prendre action pour éviter la réapparition de ces causes spéciales. On parle ici d'amélioration continuelle d'un processus qui vise à établir un processus statistiquement stable et capable de satisfaire les exigences (cibles) de la variable de réponse Y.
On vise à obtenir un processus stable et capable.

Dans la partie A on fera la mise en œuvre du SPC sur un ensemble de données. Il s'agit d'une première analyse SPC sur des données dont les résultats seront employés dans la partie B qui elle est axée sur la modélisation.

Les résultats de la partie A seront employés pour la modélisation du processus dans la partie B. Les deux méthodes de modélisation qui seront employées :

- **Régression MARS**
- **Réseaux de neurones (SANN)**

Ces méthodes de modélisation seront appliquées selon deux hypothèses :

- **Cas A : avec toutes les données sans avoir fait l'étude du SPC**
- **Cas B : avec les données retenues et identifiées par l'analyse SPC**

Ainsi, on pourra considérer, mesurer et apprécier l'impact ou non de l'utilisation du SPC.
On veut apporter une première réponse tentative à la question :

Doit-on appliquer ou non le SPC avant d'entreprendre un processus de modélisation statistique ?

MTH8302-Exer-34-partie A
étude SPC : données industrielles

Les données proviennent de l'industrie du breuvage carbonisé. Le fichier contient 37 variables dont 3 variables de réponse Y et une série de 29 variables explicatives X pour la modélisation. La période couverte est entre le 02 septembre 1999 et le 03 février 2000. Il y a un total de n = 2361 observations car on prend plusieurs observations chaque jour. Le tableau suivant présente la liste des variables ainsi que leurs définitions :

	1 Nom	2 définition / rôle
1	ID	numéro d'observation
2	Test _Time	date ceuillette observation
3	increment	facteur temps entre les observations consécutives
4	train-test	colonne pour entrainer (train) / évaluer (test) modèle
5	Brand Code	catégorie de breuvage - variable pas utilisée dans les analyses
6	Y_Carb_Volume	variable de réponse 1 : sera analysée
7	Y_Fill_Ounces	variable de réponse 2 : ne sera pas analysée
8	Y_PC_Volume	variable de réponse 2 : ne sera pas analysée
9	Carb_Pressure	variable explicative
10	Carb_Temp	variable explicative
11	PSC	variable explicative
12	PSC_Fill	variable explicative
13	PSC_CO2	variable explicative
14	Mnf_Flow	variable explicative
15	Carb_Pressure1	variable explicative
16	Fill_Pressure	variable explicative
17	Hyd_Pressure1	variable explicative
18	Hyd_Pressure2	variable explicative
19	Hyd_Pressure3	variable explicative
20	Hyd_Pressure4	variable explicative
21	Filler_Level	variable explicative
22	Filler_Speed	variable explicative
23	Temperature	variable explicative
24	Usage_cont	variable explicative
25	Carb_Flow	variable explicative
26	Density	variable explicative
27	MFR	variable explicative
28	Balling	variable explicative
29	Pressure_Vacuum	variable explicative
30	PH	variable explicative
31	Oxygen_Filler	variable explicative
32	Bowl_Setpoint	variable explicative
33	Pressure_Setpoint	variable explicative
34	Air_Pressurer	variable explicative
35	Alch_Rel	variable explicative
36	Carb_Rel	variable explicative
37	Balling_Lvl	variable explicative

Le fichier des données est BeverageManufacturing fournit en deux formats :

Statistica http://cours.polymtl.ca/mth6301/mth8302/BeverageManufacturing%20(37vX2361c).sta

Excel http://cours.polymtl.ca/mth6301/mth8302/BeverageManufacturing%20(37vX2361c).xlsx

Une copie des 5 premières observations est présentée dans ce tableau.

	1 ID	2 Test _Time	3 increment	4 train-test	5 Brand Code	6 Y_Carb_Volume	7 Y_Fill_Ounces	8 Y_PC_Volume	9 Carb_Pressure	10 Carb_Temp
1	1	1999-09-02 01:09:00	3,19		A	5,49	24,31	0,11	67,2	136,8
2	2	1999-09-02 02:09:00	4,17		A	5,39	23,95	0,23	63,2	135,0
3	3	1999-09-02 03:34:00	5,90		A	5,38	23,93	0,27	66,6	138,4
4	4	1999-09-02 06:05:00	3,33		B	5,25	23,98	0,26	64,2	140,2
5	5	1999-09-02 08:07:00	8,47		B	5,27	24,01	0,23	72,0	147,4

suite

	11 PSC	12 PSC_Fill	13 PSC_CO2	14 Mnf_Flow	15 Carb_Pressure1	16 Fill_Pressure	17 Hyd_Pressure1	18 Hyd_Pressure2	19 Hyd_Pressure3	20 Hyd_Pressure4
1	0,026	0,16	0,12	-100,00	118,4	45,8	0,00	0,00	0,00	92
2	0,042	0,22	0,08	-100,00	118,8	46,2	0,00	0,00	0,00	112
3	0,090	0,24	0,04	-100,00	119,6	45,6	0,00	0,00	0,00	116
4	0,132	0,12	0,14	-100,00	120,8	46,0	0,00	0,00	0,00	90
5	0,014	0,24	0,06	-100,00	119,8	45,2	0,00	0,00	0,00	108

suite

	21 Filler_Level	22 Filler_Speed	23 Temperature	24 Usage_cont	25 Carb_Flow	26 Density	27 MFR	28 Balling	29 Pressure_Vacuum	30 PH
1	118,6	4010	65,6	17,68	3054	1,54	722,8	3,042	-4,4	8,26
2	120,0	4012	65,6	17,60	2916	1,50	735,8	2,942	-4,4	8,26
3	120,2	4014	66,2	23,82	2948	1,52	738,8	2,992	-4,4	8,32
4	120,2	4014	65,4	18,40	2902	0,90	740,4	1,446	-4,4	8,38
5	120,8	4028	66,6	13,50	3038	0,90	692,4	1,448	-4,4	8,50

suite

31 Oxygen_Filler	32 Bowl_Setpoint	33 Pressure_Setpoint	34 Air_Pressurer	35 Alch_Rel	36 Carb_Rel	37 Balling_Lvl
0,030	120	46,0	146,2	7,14	5,44	3,04
0,030	120	46,0	147,2	7,14	5,58	3,04
0,024	120	46,0	146,6	7,16	5,44	3,02
0,064	120	46,0	147,2	6,52	5,34	1,44
0,022	120	46,0	146,2	6,54	5,34	1,38

MTH8302-Exer 34-partie A – Étude SPC (7 questions)

Lire les documents 1-2 référencés à la page2.
Les pages importantes du document2 sont : 3-14 et 20-23.

QUESTIONS

Référence : document1 (Wheeler)

Q34-A-1. Quel est l'avertissement fondamental que l'article de Wheeler essaie de communiquer ?

Q34-A-2. Quelle est votre opinion concernant ce message?

Q34-A-3. Ce message est-il connu et véhiculé dans la littérature de la science des données ?

Référence : document2 (Clément)

Q34-A-4. Quand on observe un point hors contrôle sur une carte de contrôle de processus, que doit –ton faire?

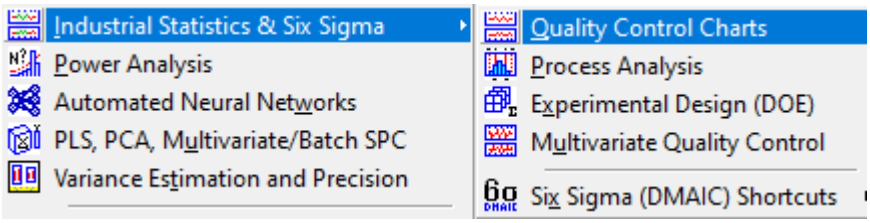
Référence : fichier de données = BeverageManufacturing – Exécution du SPC

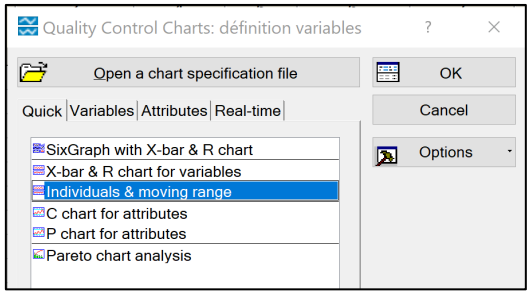
Q34-A-5. Produire une carte XmR à valeurs individuelles X et étendues mobiles mR pour la variable de réponse Y_Carb_Volume.

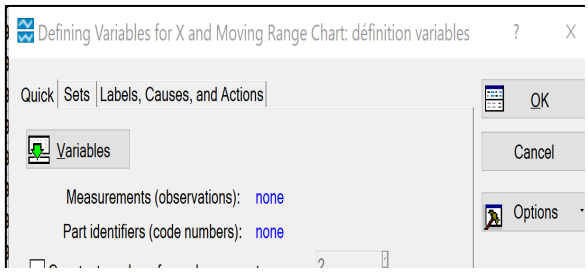
remarque : le fichier de données contient 3 variables de réponse. On aurait pu produire un carte de contrôle multidimensionnelle T^2 de Hotelling basée sur ces trois variables Y. Cela n'est pas demandé ici. Il aurait été intéressant de faire. Seule la variable Y_Carb_Volume sera étudiée dans ce devoir.

rappel : l'étendue mobile, notée mR, est la différence en valeur absolue dans une suite ordonnée (généralement par le temps) entre deux valeurs individuelles consécutives de Y.

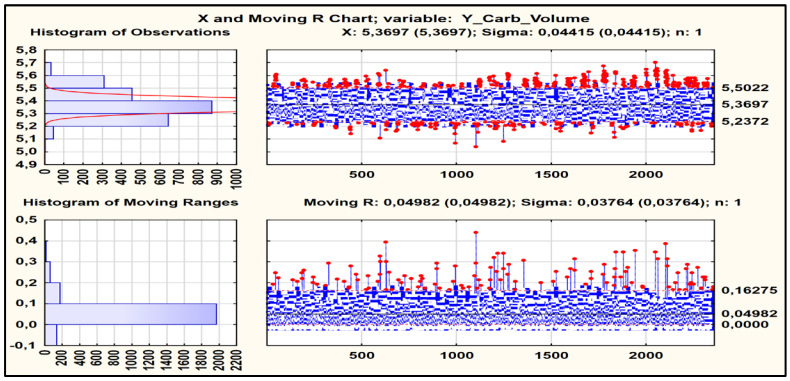
La mise en œuvre et la production du SPC se fait avec le module Quality Control Charts de Statistica







Le graphique obtenu est lourd à lire car il y a 2361 de points.
L'important pour la suite est l'identification des points hors contrôle (en rouge).
Statistica fournit le tableau de ces observations qui seront éliminées dans la suite.



Exer-34-partie A-Étude SPC (suite)

Le module du SPC contient beaucoup d'éléments de sortie. Il ne s'agit pas de tout incorporer dans le rapport mais seulement l'essentiel pour ce devoir : identifier la liste des observations qui sont hors contrôle pour la variable Y, il y en a quelque centaines. Statistica fournit aussi la liste des observations qui sont hors contrôle pour l'étendue mobile mR. Ces observations ne seront pas exclues dans la constitution du fichier (DATA2) des observations sous contrôle.

Q34-A-6. **Donner la liste des 10 premières observations qui sont déclarées hors contrôle**
Les observations seront identifiées avec leur numéro ID.

Q34-A-7. **Vous avez exécuter le SPC sur le fichier BeverageManufacturing avec la réponse Y_Carb_Volume a la question Q34-A-5.**
Constituer un fichier Statistica en retenant seulement les observations en contrôle statistique pour les valeurs individuelles. Le fichier portera le nom BeverageManufacturing (37vXnnnn).sta après élimination des observations hors contrôle. nnnn représente le nombre d'observations retenues à la suite de l'application du SPC en éliminant les observations hors contrôle.
Le nouveau fichier en question sera employé dans la deuxième partie du devoir
La variable de de réponse Y_Carb_Volume sera modalisée avec les variables explicatives X avec 2 méthodes dans la partie B.
Les modélisations de Y seront comparés dans la partie B :
avec le fichier initial (DATA1) et avec le fichier transformé (DATA2).

Exer-34-partie B -MODÉLISATION avec et sans SPC

Nous sommes maintenant en présence de 2 ensembles de données :

- **DATA1 : données initiales avec n = 2331 observations avant l'analyse SPC**
- **DATA2 : données retenues suite à l'analyse SPC sur le DATA1.**
  Le fichier DATA2 contient nnnn observations ou nnnn a été déterminé.

Nous voulons maintenant comparer les résultats de la modélisation de Y_Carb_Volume avec deux méthodes :

- **méthode1 : régression MARS**
- **méthode2 : réseaux de neurones SANN**

et avec chacun des 2 ensembles de données DATA1 et DATA2

On produira 4 modèles M1 M2 M3 M4

M1 et M3 : modélisation MARS avec DATA1 et avec DATA2
M2 et M4 : modélisation SANN avec DATA1 et avec DATA2

QUESTIONS pour la partie B

Q34-B-1 Effectuer une modélisation MARS de Y_Carb_Volume avec DATA1 (M1)
Q34-B-2 Effectuer une modélisation MARS de Y_Carb_Volume avec DATA2 (M3)
Q34-B-3 Comparer les résultats de QBB-1 et QBB-2

Q34-B-4 Effectuer la modélisation SANN de Y_Carb_Volume avec DATA1 (M2)
Q34-B-6 Effectuer la modélisation SANN de Y_Carb_Volume avec DATA2 et SANN
Q34-B-6 Comparer les résultats de QBB-4 et QBB-5

Q34-B-7 Comparer l'impact de l'opération SPC sur les résultats des analyses de modélisation en proposant votre réponse à la question

Doit-on appliquer ou non le SPC avant d'entreprendre un processus de modélisation statistique ?

RAPPORT à soumettre par courriel à bernard.clement@polymtl.ca

Transmettez avec votre rapport avec une copie des deux fichiers DATA1 et DATA2 en format Statistica (sta) ou en format Excel (xlsx) avec les noms suivants

FFFFFF_mmmmmm_DATA1.aaa **FFFFFF_mmmmmm_DATA2.aaa**

FFFFFF: nom de famille **mmmmmmm : matricule**

aaa : sta ou xlsx selon votre choix