CheMOOCs

Activités des

grains 02 et 03

Dominique Bertrand
Data-Frame, Nantes, France

Benoit Jaillais
INRA, Nantes, France

Sébastien Preys
Ondalys, Montpellier, France

Eric Latrille
INRA, Narbonne, France

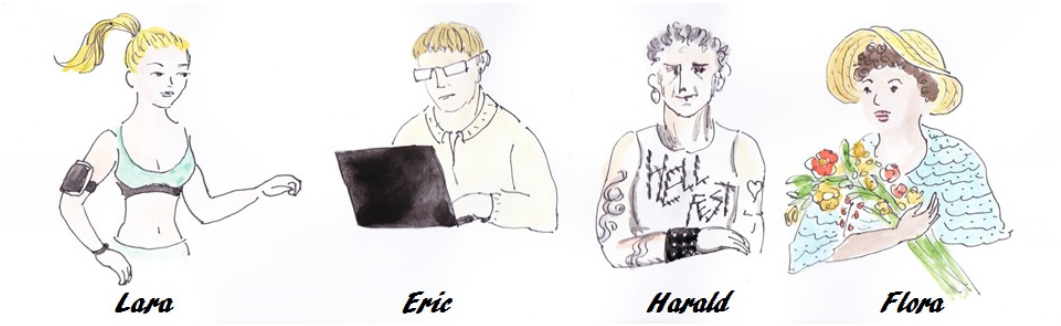
V21.09

Table des matières

Première partie

Activités du grain 02

1 Test de personnalité.



[1]

Ces sympathiques personnes ont répondu à un test de personnalité formé de 5 questions considérées comme informatives, notées de 0 (pas d'intérêt) à 10 (très important). Voici les questions posées :

— Bou : Vous intéressez vous à la Bourse ?

— Jog : Aimez-vous faire du jogging ?

— Roc : Aimez-vous la musique de style « Rock Metal » ?

— Bio : Vous intéressez-vous à l'agriculture biologique ?

— Equ : Aimez-vous monter à cheval ?

1. La figure a été dessinée par Marie Bertrand

Voici les résultats :

	Bou	*Jog*	*Roc*	*Bio*	*Equi*
Lara	0	10	2	3	6
Eric	8	4	2	2	2
Harald	0	3	10	0	0
Flora	2	3	1	9	4

FIGURE 1 – Réponses au questionnaire

Connectez-vous à ChemFlow avec votre compte personnel.

Chemflow offre la possibilité de créer des historiques, l'équivalent de répertoires dans lesquels se trouvent des données et leurs traitements. Différents historiques sont utiles pour ne pas mélanger les traitements issus de différentes données. Dans notre cas, les historiques vont être utilisés pour regrouper les données d'un grain ou de deux grains, selon la nature des exercices de la semaine de mooc.

Pour créer un nouvel historique : panneau de droite **history**, roue dentée, **create new**. Le nouvel historique apparait sous le nom *unnamed history*. Remplacer par *CheMoocs-exercice-grain02*. Ne pas oublier d'appuyer sur la touche **enter** afin de valider le changement de nom.

Nous allons maintenant entrer les données dans ChemFlow. Vous avez le choix entre deux procédures : soit les créer puis les importer vous-même, soit les récupérer dans ChemFlow.

Procédure 1 : création complète des données à partir d'un tableur.

— Les données seront reportées manuellement dans une feuille d'un tableur, Excel, OpenOffice ou LibreOffice par exemple, reproduisant le tableau 1. Attention, le séparateur décimal de votre tableur doit être le point. Si ce n'est pas le cas, reconfigurer votre tableur.

Elles seront ensuite sauvegardées au format .csv, séparateur de champ = virgule ou ',' . Par exemple, la procédure pour OpenOffice est la suivante.

enregistrer sous : choisir le répertoire de sauvegarde et le nom du fichier : *enquete* et le format : *texte csv*, plus les options *extension automatique du nom du fichier* et *éditer les paramètres du filtre*. Cliquer sur **enregistrer** puis **conserver le format actuel**. Choisir comme **séparateur de champ** la virgule (,) et effacer le contenu de **séparateur de texte**. Valider.

— Depuis ChemFlow, panneau de gauche, cliquer successivement sur **import data**, **upload file from your computer** puis **choose local file**. Sélectionner le fichier *enquete.csv* que vous venez de créer, validez avec **ouvrir**. Choisir **type**= *csv*, puis terminer l'importation avec **start**

suivi de **close**. Le fichier *enquete* apparait en jaune puis vert dans l'historique, panneau de droite. Mais il n'est pas au format csv de ChemFlow, qui utilise la tabulation comme délimiteur de champ : la dernière opération est une conversion.

— Depuis ChemFlow, panneau de gauche, cliquer sur **convert data format** puis **convert delimiters to tab**. Dans la fenêtre qui s'ouvre, et selon le choix fait précédemment, mettre une virgule dans l'option **convert all**. Vérifier que **in dataset** correspond bien à *enquete.csv* ; corriger si nécessaire. Valider avec **execute** : le fichier *enquete.csv* est rajouté à l'historique, mais cette fois il est au format tabulation, donc utilisable par ChemFlow. Vous pouvez le visualiser en cliquant sur l'œil. Notez que le nom n'a pas changé ! Pour ne pas vous tromper, renommez-le : cliquez sur le crayon puis dans **edit attributes/name** mettre *enquete.tab* et cliquez sur **save** pour enregistrer.

Procédure 2 : récupération des données dans ChemFlow.

— Les données sont disponibles dans : **shared data / data libraries / chemoocs / grain 02 / enquete.tab**. Sélectionnez le fichier en cochant la case à gauche du nom, puis cliquez sur **to history**. Le fichier *enquete.tab* devient visible dans votre historique : si vous cliquez sur l'oeil, vous aurez le tableau 1 ci-dessus.

Questions

— 1.1. Si on appelle $\mathbf{X}$ la matrice des résultats, que représente $\mathbf{X}(3, 2)$?

— 1.2. Calculez la matrice des distances Euclidiennes des réponses des différents participants. Dans ChemFlow, panneau de gauche, cliquez successivement sur **statistics** puis **matrix distance**. Vérifiez que le champ **select X data** contient bien *enquete.tab* et que **distance choice** contient bien *euclidian*, corriger si nécessaire en cliquant sur les noms affichés et sélectionner les bonnes valeurs. Puis cliquer sur **execute**. Le résultat apparait dans l'historique, il est visualisé à l'écran en cliquant sur l'œil.

  — Quelles sont les personnes les plus proches ?

  — Quelles sont les personnes les plus éloignées ?

— 1.3. Calculez la norme du vecteur des réponses de Harald. Ce calcul sera fait manuellement, avec une calculatrice.

2 Spectres visible - proche infrarouge de farines de blé.

[2]

Chargez dans votre historique les fichier *x_140farines.tab* et *y_140farines.tab* qui se trouvent aussi dans **shared data / data libraries / chemoocs / grain 02** . Ces deux fichiers contiennent respectivement les spectres de 140 farines de blé pour le premier, les teneurs en protéines et la nature dure/tendre du blé pour le second.

Questions

— 2.1. Dessinez la courbe du spectre moyen, puis dessinez la courbe des écart-types à toutes les longueurs d'onde. Que peut-on conclure de ces deux courbes ?

Dans ChemFlow, le spectre moyen est obtenu depuis **statistics** par la fonction **mean** ; il s'appelle *mean on x_140farines.tab*. De même, le spectre des écarts-types est obtenu depuis **statistics** par la fonction **standard deviation** et il s'appelle *sd on x_140farines.tab*.

Pour éditer la figure, deux options sont proposées : soit utiliser **spectra plot**, une fonction dédiée à la visualisation de spectres, soit utiliser **scatter**, une fonction plus généraliste qu'on appliquera ici à des spectres.

— Utilisation de **spectra plot :**

Renseigner les différents champs :

— **plot title** → *spectre moyen*

— **label for x axis** → *longueurs d onde*

— **label for y axis** → *absorbances*

— **spectra/dataset** → *mean on x_140farines.tab*

2. Photo INRA

Laissez les autres options par défaut et validez avec **execute**.

— Utilisation de **scatter plot :**

Attention, **scatter plot** utilise des colonnes de chiffres, et ici nos données de moyenne et d'écart-type sortent en ligne. Il faut donc transposer les vecteurs moyenne et écart-type afin de les disposer en colonne, avec la fonction **utils/transpose matrix**.

Après avoir transposé, renseigner les différents champs :

— **plot title** → *spectre moyen*

— **label for x axis** → *longueurs d onde*

— **label for y axis** → *absorbances*

— **plot type** → *line/multi lines*

— **dataset** → *trans(mean on x_140farines.tab)*

— **column for x axis** → *c1*

— **column(s) for y axis** → *c2 :x*

— **line type** → *solid*

Laisser les autres options par défaut et validez avec **execute**.

Activer **scratchbook**, en haut de l'écran, afin de voir les figures produites. Cliquer sur l'œil pour afficher la figure. Vous pouvez télécharger la figure au format pdf en cliquant sur la flèche. Recommencez avec le fichier *sd on x_140farines.tab* et en mettant *ecarts-types des variables spectrales* dans **plot title**.

— 2.2. Dessinez le corrélogramme entre les spectres et la teneur en protéines. Que peut-on en conclure ?

Aller dans **plots** puis **correlogram**. Renseigner les options suivantes :

— **dataset containing the spectra** → *x_140farines.tab*

— **dataset containing the variables** → *y_140farines.tab*

— **variable to correlate with spectra** → *c2 :protref*

— **plot title** → *correlogramme*

— **label for x axis** → *longueurs d'onde*

puis **execute**.

— 2.3. Les échantillons de blé sont soit des blés durs, soit des blés tendres. Construire l'histogramme de la variable « protéines » en représentant l'appartenance des échantillons à leur classe. Que peut-on en conclure ?

L'histogramme est obtenu depuis **plot** avec la fonction **histogram**. Dans **dataset** choisir

y_140farines.tab. La valeur de **bin width** (le pas qui sépare les classes) peut être laissé à la valeur par défaut, soit 1. Dans **use a column of a dataset as bar color** prendre l'option **yes** ce qui ouvre deux champs. Pour **dataset** sélectionner le fichier *y_140farines.tab* et pour **column factor choice color bar** sélectionner *c3 :dur/tendre*.

— 2.4. Construire de la même manière l'histogramme de la longueur d'onde *2000 nanomètres*. La figure finale devrait comporter dans les 10 - 20 classes. Que peut-on en conclure ?
L'histogramme est construit comme précédemment, avec le fichier *x_140farines.tab*. Depuis le dernier histogramme de l'historique, vous pouvez utiliser l'icône recyclage, les 2 flêches inversées : cela rappelle la fonction avec les mêmes arguments, seuls un petit nombre est à changer. Choisir la bonne longueur d'onde dans **column for x-axis**. Ajuster la valeur de **bin width** par tatonnements, diminuer la valeur pour augmenter le nombre de classes, et inversement.

Deuxième partie

Activités du grain 03.

3 Informations générales

3.1 Pré-requis.

— vidéo + pdf du grain 03 ;
— tutoriel ChemFlow.

3.2 Durée estimée.

— 30 minutes.

3.3 Description du jeu de données.

Les données ont été produites par l'Université d'Aix-Marseille, équipe de Nathalie Dupuy. Des analyses chimiques ont été réalisées sur 187 huiles d'olives dont l'origine géographique était connue. Les données se composent de :

— un jeu comportant en colonne les analyses biochimiques de 14 acides gras et du squalène sur ces 187 échantillons disposés en ligne : fichier *ags.tab* dans ChemFlow.

— un codage des 187 échantillons selon les 6 origines géographiques : AP=Aix en Provence, HP=Haute Provence, NI=Nice, NM=Nimes, NY=Nyons, VB=Vallée des Baux de Provence ; fichier *origine.tab* dans ChemFlow.

3.4 Chargement des données dans ChemFlow.

— Dans ChemFlow, créer un nouvel historique, nommé : *CheMoocs-exercice-grain03* ;

— importer toutes les données disponibles en les sélectionnant puis en cliquant sur l'icône **to history** qui se situe en haut au centre ;

— retourner à la page d'accueil (cliquer sur **galaxy/chemflow**) et vérifier que les données ont bien été importées.

4 Question 1 : visualisation des données.

— Visualiser les données d'origines géographiques sous forme d'un tableau.

Activer **scratchbook** et cliquer sur l'œil au niveau des données dans le panneau de droite.

— Visualiser les premières lignes des données d'analyse biochimique sous forme d'un tableau de données.

Utiliser **scratchbook** et l'œil. Puis revenir à la page d'accueil.

— Visualiser les données sous forme de boîtes à moustaches ou de graphes de dispersion - ce type de figure montre la moyenne de chaque variable, et aussi sa dispersion autour de sa moyenne .

Panneau de gauche : **plot/boxplot/dataset** et choisir *ags.tab*. Sélectionner toutes les variables, soient les 15 colonnes du fichier.

— Que peut-on dire de la dispersion des données ? Qu'en déduisez-vous pour le choix des options de centrage et normalisation d'une analyse en composantes principales ?

5 Question 2 : réalisation et visualisation d'une ACP.

— Réaliser une ACP sur le jeu de données *ags.tab* avec centrage et normalisation préalable.

  — **exploration/pca/centering** → *yes*

  — **exploration/pca/scaling** → *yes*

— **exploration/pca/compute outliers statistics** → *no*

— Visualiser les valeurs numériques du pourcentage de variance expliquée en fonction du nombre de composantes principales.

Utiliser l'œil au niveau des données de l'historique : *pca on ags.tab : explained variance (%)*

— Tracer le diagramme de variance expliquée (= éboulis). Qu'observe-t-on ?

Deux possibilités d'édition : soit un graphique en barres, soit une courbe. Attention à ne pas utiliser de caractères accentués.

Graphique en barres :

Utiliser **plot/barplot**. Renseigner :

— **dataset** → *pca on ags.tab : explained variance (%)*

— **column for x axis** → *c2 : explained variance*

— **label for y axis** → *pourcentage de variance*

Courbe :

Utiliser **plot/scatterplot**. Renseigner :

— **plot title** → *eboulis des valeurs propres*

— **label for x axis** → *composantes principales*

— **label for y axis** → *pourcentage de variance*

— **plot type** → *lines and points*

— **dataset** → *pca on ags.tab : explained variance (%) :*

— **column for x axis** → *c1* ;

— **column for y axis** → *c2 :explained variance*

Finir par **execute**.

— Tracer les plans 1-2 et 1-3 de la carte factorielle des individus (score plot) avec des couleurs différentes pour chaque origine. Observe-t-on des outliers, si oui lesquels ? Peut-on différencier les différentes origines à l'aide de cette visualisation ?

Les scores sont dans l'historique : *pca on ags.tab : scores.*

Il faut utiliser **plot/scatter plot/** :

— **plot/scatter plot/label for x axis** → *pc 1*

— **plot/scatter plot/label for y axis** → *pc 2*

— **plot/scatter plot/series/plot type** → *points*

— **plot/scatter plot/series/plot type/x-dataset** → *pca on ags.tab : scores*

— **plot/scatter plot/series/plot type/column for x axis** → *c2 : pc1*

- **plot/scatter plot/series/plot type/y-dataset** → *pca on ags.tab :scores*
- **plot/scatter plot/series/plot type/column for y axis**→ *c3 : pc2*
- **plot/scatter plot/series/plot type/add first column on x-dataset as sample label**→ *yes*
- **plot/scatter plot/series/plot type/use a column of a dataset as point color**→ *yes*
- **plot/scatter plot/series/plot type/use first column as sample label/dataset** → *origine.tab*
- **plot/scatter plot/series/plot type/use first column as sample label/column for color** →*c2 : origine*

Faire de même pour PC1 et PC3.

6 Question 3 : interprétation des résultats d'une ACP.

- Tracer les plans 1-2 et 1-3 du cercle des corrélations des variables. Peut-on identifier des groupes de variables identiques ? Lesquels ?
  - **plots/correlation circle/dataset containing the pca scores** → *pca on ags.tab : scores*
  - **plot/correlation circle/column for x axis** → *c2 :pc1*
  - **plot/correlation circle/column for y axis** → *c3 :pc2*
  - **plot/correlation circle/dataset containing the variables to project on the circle** → *ags.tab*
  - **plot/correlation circle/variables to project on the circle** → *select all*
  - **plot/correlation circle/plot title** → *cercle des correlations, pc1-pc2*
- Proposez une interprétation des axes factoriels (les composantes principales).