Projets en groupe — printemps 2025

Benjamin Quost

21/03/2025

Résumé

Vous avez un projet à réaliser au cours de l'UV, en groupes, qui a une double vocation pédagogique et d'évaluation. Il a pour objectif d'appliquer les méthodes étudiées au cours du semestre sur des jeux de données réelles. Vous rendrez compte du travail effectué et de la démarche adoptée lors d'une soutenance sur la base d'un poster ; vous vous appuierez également sur un bref rapport, qui ne sera pas évalué mais vous permettra de donner des précisions sur votre travail.

Ce document précise les modalités du projet (conditions de réalisation et d'évaluation), donne quelques conseils généraux pour le bon déroulement de votre travail, et fournit quelques indications sur la rédaction du rapport. Celui-ci, *qui n'excèdera pas quatre pages* (en style « double-colonne »), sera rédigé en LaTeX. Vous pouvez réutiliser le fichier source du présent document. Un premier rendu intermédiaire sera fait peu avant la revue de projet à mi-parcours. La version finale sera rendue avant les soutenances.

Introduction

Ce document présente les conditions de réalisation du projet de SY09 pour le semestre de printemps 2025.

Le document présente tout d'abord (en partie 1) les modalités de réalisation du projet, les livrables à fournir, et les conditions d'évaluation. Le travail sera réalisé en groupes ; il portera sur l'analyse et le traitement d'un jeu de données, *choisi avec soin* par le groupe. Le travail réalisé sera restitué lors d'une soutenance en fin de semestre, sur la base d'un poster ; des éléments complémentaires pourront être donnés dans un rapport.

Ce rapport fera l'objet de *deux rendus* : un premier avant la revue de projet à mi-parcours, et un rendu final avant les soutenances de projets. Il ne sera pas évalué, mais vous permettra de donner des compléments d'information sur votre travail : il est donc particulièrement important pour la mise en valeur de ce dernier. Pour cette raison, ce document donne (en partie 2) quelques conseils pour la rédaction de rapports de qualité.

1 Modalités du projet

Cette première partie donne quelques précisions sur les conditions de réalisation du projet et l'évaluation du travail réalisé par les groupes.

1.1 Groupes et choix des données

Le travail sera réalisé en groupes de trois personnes, tirés au hasard. Chaque groupe travaillera tout au long du semestre sur un même jeu de données *choisi avec soin* sur Kaggle. Deux groupes ne peuvent pas choisir le même jeu de données. Nous mettrons en place une activité Moodle permettant de déclarer les données choisies, et vérifierons qu'il n'y a pas de doublons.

Nous recommandons de veiller aux points suivants lorsque vous choisirez votre jeu de données :

- bien prêter attention aux types de traitements auxquels les données se prêtent : il faut pouvoir leur appliquer des méthodes de visualisation, de clustering, et de classification supervisée [1] ;
- éviter les données temporelles, les données textuelles (pas d'analyse de séries temporelles ou de traitement du langage en SY09) ; les jeux de données d'images sont souvent difficiles à exploiter.

Il ne vous est pas demandé d'utiliser des méthodes qui ne sont pas étudiées dans le cadre de l'UV.

1.2 Soutenances

Le travail réalisé sera restitué par le groupe lors d'une soutenance, sur la base d'un poster — nous donnerons davantage d'informations sur la réalisation de ce poster dans le courant du semestre. Seule la prestation orale, évaluée lors d'une soutenance entre le 16/06/2025 et le 21/06/2025, sera notée. Néanmoins, le temps imparti à chaque groupe étant limité, chaque groupe pourra s'appuyer sur un rapport, qui constituera un complément d'information sur le travail réalisé.

1. Éviter les données adaptées à un unique type de traitement (ex. données étiquetés « visualisation ») ; éviter les jeux de données « régression », thématique peu traitée en SY09.

Soulignons qu'il est important que chaque membre d'un groupe ait une *maîtrise raisonnable de l'ensemble des travaux réalisés* par le groupe et présentés lors de la soutenance. Nous veillerons à évaluer ce point lors des soutenances, en posant des questions ciblées à chacun, de manière à évaluer son niveau d'implication et à juger de sa maîtrise des notions étudiées en SY09.

1.3 Revue de projet

Une première version du rapport, ne rendant compte que des travaux réalisés lors de la première partie du semestre, sera rendue avant la revue de projet à mi-parcours, qui aura lieu entre le 21/04/2025 et le 26/04/2025. Cette version intermédiaire sera évaluée par plusieurs étudiants, dans le même esprit que les travaux de recherche présentés dans le cadre d'une conférence : chaque groupe verra son travail évalué par plusieurs pairs (autres étudiants), et chaque étudiant devra évaluer au moins un rapport intermédiaire.

Les rapporteurs de chaque travail seront choisis au hasard, de manière anonyme, et auront à cœur de *critiquer (de manière constructive)* le travail réalisé, en posant des questions et suggérant des pistes d'analyse. Le but est d'aider un groupe à progresser dans l'analyse des données choisies, et de se rendre compte (en analysant les travaux faits par d'autres) des limitations de son propre travail. Les intervenants synthétiseront les analyses faites par les étudiants, et pourront de même faire un retour sur ces analyses.

Nous donnerons davantage de détails sur les délais et la manière dont les rapports intermédiaires seront mis à disposition pour être relus et commentés lors de la revue à mi-parcours (ainsi que pour le rendu de la version finale du rapport à la fin du semestre). Nous préciserons à cette occasion comment chacun pourra transmettre son évaluation des travaux qui lui auront été assignés, et comment prendre ensuite connaissance des retours sur les commentaires faits dans le cadre de ses évaluations.

2 Rapports

Cette partie donne quelques conseils de réalisation des rapports de projets. Ces derniers ne seront pas évalués, mais permettent de communiquer des éléments complémentaires qu'il serait difficile de développer lors de la soutenance et de mettre en valeur les travaux réalisés : il est donc important d'en soigner la réalisation.

2.1 Introduction

La rédaction d'un rapport est toujours un exercice délicat : le document doit être concis, c'est-à-dire suffisamment synthétique sans pour autant omettre d'informations importantes. Il devra être rédigé avec LaTeX, dans le style du document présent (dont vous pourrez reprendre le code source), et ne devra pas excéder cinq pages. Il sera mis en ligne sur la page Moodle de l'UV SY09.

Le document pourra respecter la structure suivante :

1. une introduction, rappelant la problématique (présentation brève des données, questions auxquelles le travail cherche à répondre, plan) ;
2. une partie principale, articulée suivant la problématique d'étude, pouvant contenir différentes parties : présentation des données, puis des différentes analyses effectuées ;
3. une conclusion, résumant les résultats principaux et présentant d'éventuelles perspectives dont la mise en œuvre dépasse le cadre du TP.
4. D'éventuelles annexes pourront contenir le code, les figures (si elles sont trop nombreuses pour être mises dans le texte), voire des détails de calculs un peu longs.

Nous donnons ci-dessous quelques conseils généraux sur le contenu d'un rapport (partie 2.2) et sur sa forme (partie 2.3) : en particulier, la restitution de formules mathématiques est abordée au paragraphe 2.3.2, de code au paragraphe 2.3.3, et de tableaux ou de figures au paragraphe 2.3.4. La question de la bibliographie est brièvement traitée au paragraphe 2.4.

Remarque 1 (LaTeX) *Nous imposons l'usage de LaTeX, car cette solution logicielle riche, performante, libre et gratuite, permet d'élaborer des documents de qualité. L'usage de LaTeX ne garantit pas la qualité du document final (le respect des bonnes pratiques de rédaction revient aux auteurs d'un document), mais la favorise néanmoins : en effet,*

- *l'utilisation de balises incite à dissocier le contenu de la mise en page, qui est très facile à modifier (comme par exemple le choix du style « double colonne » dans l'en-tête du document) ;*
- *cela permet une grande flexibilité dans la manipulation du contenu (paragraphes, figures, tableaux, algorithmes, équations, etc), même en modifiant la structure du rapport ;*
- *la gestion des équations (et des écritures mathématiques en général), ainsi que du code, est particulièrement aisée.*

La mise en page est rigoureuse, et respecte par défaut les règles d'usage en typographie : on en tire donc en général de beaux documents, plus faciles à lire.

Le prix à payer est d'investir un peu d'énergie pour découvrir LaTeX. Ce n'est pas si difficile, et le retour sur investissement est rapide et conséquent. L'existence d'une documentation très abondante et de fichiers types rend la tâche encore plus aisée. Un manuel très complet est disponible en ligne [3] (une version française, plus ancienne, est disponible par exemple à cette URL).

2.2 Contenu du rapport

L'analyse des résultats obtenus lors de votre travail sera effectuée lors de la soutenance. Étant donné le peu temps disponible pour cette présentation orale, il vous est recommandé de détailler votre travail dans un rapport que vous pourrez utiliser comme support. Ce rapport devant être succinct (cinq pages double colonne), il devra en conséquence être soigneusement élaboré pour avoir une réelle valeur ajoutée.

Le rapport vous permettra de préciser les choix faits lors du traitement du jeu de données que vous avez choisi, et de préciser le cheminement intellectuel qui vous a conduit aux résultats présentés lors de la soutenance. Les méthodes ou algorithmes utilisés peuvent être décrits succinctement. Le code source devrait être évité dans le rapport, à moins qu'ils ne présente un intérêt particulier ; il peut être consigné en annexe, pour ne pas perturber la lecture.

2.3 Forme

Nous donnons ci-dessous quelques conseils généraux sur la forme du rapport, concernant la mise en page et le choix des polices, la présentation de résultats formels (mathématiques), techniques (algorithmes), ou graphiques (tableaux, figures).

2.3.1 Police et mise en page

Un bon document doit être sobre : les fioritures fatiguent à long terme le lecteur, en particulier lorsqu'il s'agit de lire un grand nombre de documents en un temps limité. Pour cette raison, n'abusez pas d'effets de style. On évitera l'usage de caractères **gras** ou d'une police soulignée : pour mettre en valeur un élément de texte, *on préfèrera utiliser l'italique*. Les changements de taille ou de style de police nuisent à l'harmonie graphique du document, et sont donc à proscrire. On pourra faire une exception à cette règle dans des cadres bien spécifiques, par exemple en utilisant le gras pour distinguer les vecteurs des variables, ou une police particulière pour écrire du code (paragraphes 2.3.2 et 2.3.3).

Un document rédigé en pleine page (« simple colonne ») est plus agréable à lire que présenté sur deux colonnes, mais prend plus de place : si le nombre de pages est contraint, le style « double colonne » est préférable. Il n'est pas besoin de rajouter des sauts de ligne ou des espaces, LaTeX fait ça très bien tout seul. N'abusez pas des notes de bas de page, qui fractionnent la lecture [2] ; les précisions doivent rester dans le corps du texte (par exemple entre parenthèses), ou être développées dans un paragraphe à part.

2.3.2 Mathématiques

L'un des nombreux avantages de LaTeX est l'aisance avec laquelle on peut écrire et gérer les formules mathématiques. Vous pouvez ainsi insérer des équations seules, comme c'est le cas pour l'équation (1) :

$$f(x) = ax^2 + bx + c, \tag{1}$$

ou des suites d'équations (alignées, pour plus de clarté) comme les équations (2)-(3) :

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top A\boldsymbol{x} + b^\top \boldsymbol{x} + c, \tag{2}$$

$$\text{t.q. } \boldsymbol{x} \in \mathbb{R}^2, \tag{3}$$

où A et b sont définis par :

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Remarquons que cette dernière équation n'a pas été étiquetée : on ne peut pas y faire référence (sauf avec une périphrase, ce qui nuit à la concision du propos). Une équation longue (comme l'équation (4)) peut figurer sur plusieurs lignes :

$$f(x_1, x_2) = a_{11}x_1^2 + a_{22}x_2^2 + a_{21}x_1x_2 + a_{12}x_1x_2 + b_1x_1 + b_2x_2. \tag{4}$$

Les développements mathématiques complexes ou d'intérêt secondaire doivent autant que possible figurer dans un paragraphe à part ou en annexe, pour ne pas gêner la lecture (voir par exemple l'annexe A).

Remarquons que LaTeX offre la possibilité de définir des commandes pour les usages répétés, comme par exemple pour les symboles « $\boldsymbol{x}$ » et « $^\top$ » (voir l'en-tête du fichier source de ce document). Pour plus de renseignements, on pourra se reporter à [3] (ou à la version française, plus ancienne).

2. Les notes de bas de page sont plutôt dévolues aux apartés ; mal utilisées, elles gênent la lecture en en brisant le fil.

2.3.3 Algorithmes et code source

Comme nous l'avons dit plus haut, le corps du texte ne devrait pas comporter de code source complexe, comme des fonctions ou des scripts : cela nuit au confort de lecture et à la clarté du document. La mention d'un petit nombre de commandes, si nécessaire, est tolérable. Le code plus complexe peut être consigné en annexe.

Nous avons défini, dans l'en-tête de ce document, la commande `\pycode` pour identifier du code `python` dans le corps du texte, par exemple pour faire référence à une fonction générique (comme `numpy.mean`). La retranscription de code source (qui devrait sauf exception être faite en annexe du rapport) nécessitera l'usage d'un environnement. Un exemple est donné dans l'annexe B.

Enfin, si vous avez développé un algorithme particulier pour effectuer un traitement spécifique sur des données, il vous est possible de le présenter de manière formelle et d'y faire référence grâce à un environnement adapté, comme celui fourni par le package LaTeX `algorithm2e`.

2.3.4 Tableaux et figures

D'une manière générale, les informations communiquées dans le rapport doivent être lisibles et informatives. Les tableaux et les figures peuvent vous y aider. Nous rappelons ici quelques règles de bon sens.

LaTeX permet de créer des tableaux et des figures assez facilement, pour ensuite y faire référence dans votre rapport. Vous trouverez des exemples dans ce document, avec le tableau 1 et les figures 1 et 2. La disposition des tableaux et figures (appelés « objets flottants ») est le point qui peut être délicat lors de la rédaction d'un document : la compilation peut les positionner dans des endroits peu désirables. Il est parfois nécessaire de tâtonner en positionnant le code correspondant aux objets flottants à différents endroits dans le code source du document (et en recompilant à chaque fois) jusqu'à obtenir une configuration satisfaisante.

Table 1 – Exemple de tableau.

UV	niveau	remarque
SY02	branche	classe
SY09	filière	top
SY19	filière	très bien

On évitera d'afficher trop de tableaux ou de figures. Ceux qui seront choisis devront être optimisés, de manière à communiquer autant d'informations *pertinentes* que possible en un minimum de place. Il faut éviter les tableaux ou figures « orphelins » (sans légende, pas numérotés, sans référence). La légende, synthétique (le texte doit contenir l'essentiel de l'information), doit permettre au lecteur de situer la figure par rapport au corps du texte. Notons que par convention, la légende se met en haut pour un tableau et en bas pour une figure.

On apportera une attention particulière à la lisibilité : éviter les valeurs avec trop de chiffres significatifs, utiliser un style épuré (le tableau 1, par exemple, est défini avec un minimum de lignes séparatrices). On n'oubliera pas d'indiquer les échelles sur les graphiques, d'étiqueter les axes, etc. Les couleurs et les symboles permettent d'ajouter une information supplémentaire parfois précieuse (par exemple, sur la figure 1, elle représente l'espèce d'iris dans le célèbre jeu de données collecté par Anderson [1] et popularisé par Fisher [2]). Les documents étant généralement imprimés en noir et blanc, on utilisera des symboles en plus des couleurs.

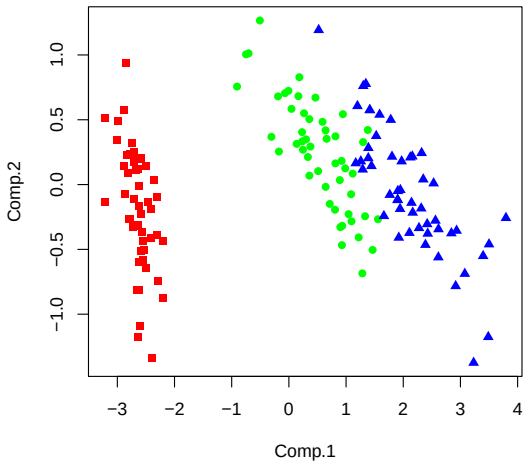

Figure 1 – Iris de Fisher, premier plan factoriel.

Comme nous l'avons écrit ci-dessus, on évitera les tableaux ou graphiques peu informatifs ou sans intérêt, qui polluent le document. Par exemple, pour comparer différentes populations en utilisant des diagrammes en boîte (*boxplots*), on les juxtaposera dans la même figure. L'affichage des valeurs prises par une variable en fonction de l'indice des individus dans le jeu de données (figure 2), erreur trop souvent rencontrée dans les rapports, est typique de ce qu'il faut éviter.

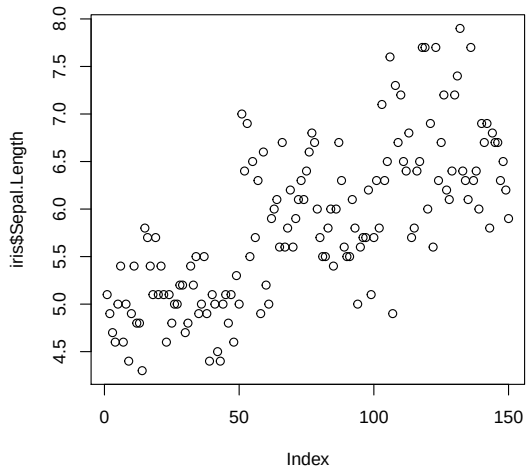

FIGURE 2 – Figure pas optimisée : peu informative (abscisse sans intérêt), moche, sans couleur ; à proscrire.

2.4 Bibliographie

C'est une partie parfois négligée par les étudiants, alors qu'elle est d'une importance fondamentale, puisqu'elle touche à la valorisation du travail et au problème du plagiat. L'intérêt de la bibliographie est de faire clairement référence aux connaissances antérieures (articles, livres, sites web, projets, code existant) à partir desquelles un travail est mené, et permet de positionner ce dernier par rapport à l'existant. Sans bibliographie, les contributions ne sont pas mises en valeur ; et la paternité du travail présenté dans le document peut être remise en cause.

Nous ne traiterons pas la question de la bibliographie en détails. Il existe en LaTeXplusieurs possibilités. La solution la plus courante (et utilisée dans ce document), BibTeX, nécessite de renseigner les références par type, dans un fichier à part (extension `.bib`). La compilation du document se fait en plusieurs étapes. À chaque ajout d'une référence, il faudra ainsi compiler une fois avec LaTeX, puis une fois avec BibTeX, et enfin à nouveau une (voire deux) fois avec LaTeX.

2.5 Conclusion

Ces conseils ont pour objectif de vous permettre de rédiger des documents plus agréables et faciles à lire. L'utilisation d'un logiciel tel que LaTeX vous poussera à rédiger un compte-rendu avec rigueur, en sélectionnant et en structurant l'information que vous souhaitez restituer. Cette exigence n'est pas une perte de temps : elle sera profitable à la qualité de votre document, qui sera plus agréable à lire et à reprendre par la suite.

Ce document ne dispense que des conseils élémentaires de rédaction et de mise en forme, et laisse de côté certains aspects jugés peu importants pour un rapport de projet. Il existe de nombreux documents permettant d'aller plus loin avec LaTeX ou décrivant ses extensions (graphiques avec `tikz`, diapositives avec `Beamer`, etc) : la littérature disponible est bien plus abondante que l'introduction générale [3].

A Éléments sur la loi normale

Voici un premier exemple d'appendice avec une définition (paragraphe A.1) de la loi normale multivariée et une propriété de cette distribution relative au conditionnement (paragraphe A.2).

A.1 Définition

La fonction de densité d'un vecteur aléatoire $\boldsymbol{X}$ distribué suivant une loi normale multivariée d'espérance μ et de matrice de covariance Σ est [4]

$$f_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left((\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (5)$$

où $\boldsymbol{\mu}$ est un vecteur réel de dimensions $p \times 1$ et $\boldsymbol{\Sigma}$ une matrice symétrique définie-positive de dimensions $p \times p$.

A.2 Propriété

Propriété 1 (Loi normale et conditionnement) *Soit $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ un vecteur aléatoire gaussien (voir paragraphe A.1). Supposons que $\boldsymbol{X}$ peut être séparé en deux sous-vecteurs $\boldsymbol{X}_A$ et $\boldsymbol{X}_B$, où A et B indiquent les indices des variables correspondantes :*

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_A \\ \boldsymbol{X}_B \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix}.$$

Alors $\boldsymbol{X}_A | \boldsymbol{X}_B = \boldsymbol{x}_B$ est un vecteur aléatoire gaussien :

$$\boldsymbol{X}_A | \boldsymbol{X}_B = \boldsymbol{x}_B \sim \mathcal{N}(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B}), \quad (6)$$

où

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\boldsymbol{x}_B - \boldsymbol{\mu}_B),$$
$$\boldsymbol{\Sigma}_{A|B} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}.$$

B Code

Nous reproduisons ci-dessous du code permettant de charger les données `Iris` et de représenter les boxplots (diagrammes en boîte) obtenus pour chaque variable quantitative, au moyen de la bibliothèque `seaborn` :

```
import seaborn as sns

iris = sns.load_dataset("iris")
iris.drop(columns=["species"]).boxplot()
```

Le graphique obtenu est représenté dans la figure 3, qui permet ainsi de comparer les distributions empiriques de ces quatre variables descriptives.

Pour afficher le code en couleurs, LaTeX utilise l'environnement `minted`, qui nécessite d'installer le package `pygments`, et de compiler le document avec l'option `--shell-escape` :

```
pdflatex --file-line-error --synctex=1
  ↪  --shell-escape projet.tex
```

En cas de difficultés, vous pouvez appeler à l'aide sur le forum de la page Moodle de SY09.

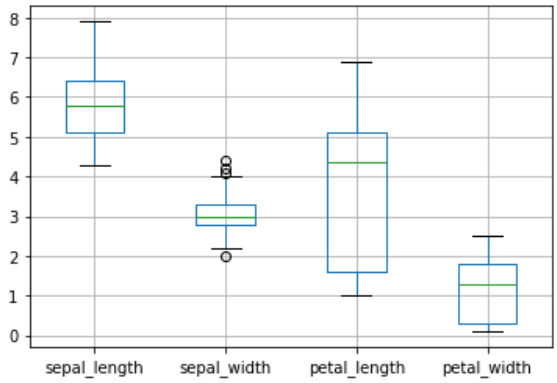

FIGURE 3 – Boxplots des quatre variables descriptives des données `Iris`.

Références

[1] E. Anderson. The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59 :2–5, 1935.

[2] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2) :179–188, 1936.

[3] T. Œtiker, H. Partl, I. Hyna, and E. Schlegl. *The Not So Short Introduction to LaTeX 2ε*, 2021.

[4] K. B. Petersen and M. S. Pedersen. The Matrix Cookbook. Technical report, Technical University of Denmark, 2012.