

ETUDE DE CAS EN STATISTIQUE ET INFORMATIQUE DÉCISIONNELLE : UN EXEMPLE BASE SUR UNE ENQUETE EN DUT STID

Frédérique Letué¹ & Marlène Villanova-Oliver²

¹ STID/IUT2/UPMF et LJK, Frederique.Letue@iut2.upmf-grenoble.fr
² STID/IUT2/UPMF et LIG, Marlene.Villanova-Oliver@iut2.upmf-grenoble.fr

Résumé. Dans le nouveau Programme Pédagogique National (PPN) du DUT STID figure un module intitulé « étude de cas en statistique et informatique décisionnelle ». Si les projets tuteurés par exemple constituent depuis toujours dans la formation une occasion d’associer au sein d’un même cadre pédagogique la statistique et l’informatique, il s’agit ici d’instaurer un enseignement alliant explicitement les deux disciplines. Nous présentons la première version de ce module, tel qu’il a eu lieu à l’automne 2014 au sein du Département STID de Grenoble. Ce module s’adresse aux étudiants de 2^{ème} année de DUT et s’est tenu au cours du premier semestre. Nous détaillons le contexte dans lequel s’inscrit le module en faisant le point sur les cours déjà dispensés à ce public en statistique et informatique au moment où a eu lieu le module. Celui-ci est ensuite plus particulièrement présenté à travers son format, le cas sur lequel il s’appuie et son contenu. Nous mettons ensuite en regard les attentes du cours et les difficultés rencontrées permettant ainsi de dégager des pistes d’amélioration possibles.

Mots-clés. Enseignement de la statistique, étude de cas.

Abstract. In the new “Programme Pédagogique National (PPN)” of the “DUT STID” stands a new course unit entitled “case study in statistics and business intelligence”. Until now, the “projets tuteurés” gave students the opportunity to link statistics and computer science in a same framework. Here, a course is explicitly dedicated to the association of both fields. We present the first version of this course as it was given in Grenoble in autumn 2014, during the first semester. This unit is intended for the second year DUT students. We detail the context of the course, listing the courses already given to this student audience in statistics and computer science. Its format, the case we deal with and its content are presented. We finally highlight the course expectations, the difficulties encountered by the students, and improvements tracks.

Keywords. Statistics teaching, case study.

1 Introduction

Dans le nouveau Programme Pédagogique National (PPN) du DUT STID [1] figure un module intitulé « étude de cas en statistique et informatique décisionnelle ». Si les projets tuteurés par exemple constituent depuis toujours dans la formation une occasion d’allier au sein d’un même cadre pédagogique la statistique et l’informatique, il s’agit ici d’instaurer un enseignement alliant explicitement les deux disciplines. Nous présentons ici la première version de ce module, tel qu’il a eu lieu à l’automne 2014 au sein du Département STID de Grenoble. Nous détaillons le contexte dans lequel s’inscrit le module et le public auquel il s’adresse en faisant également le point sur les cours déjà dispensés à ce public en statistique

et informatique au moment où a eu lieu le module. Celui-ci est ensuite plus particulièrement présenté à travers son format, le cas sur lequel il s'appuie et son contenu. Nous mettons ensuite en regard les attentes du cours et les difficultés rencontrées pour établir les pistes d'amélioration pour les années à venir.

2 Contexte

La version 2013 du PPN STID est notamment marquée par la place faite à de nouveaux enseignements en informatique décisionnelle en deuxième année de DUT. De ce fait, plusieurs enseignements visant à asseoir des compétences requises pour appréhender l'informatique décisionnelle ont été avancés en première année. Le module « étude de cas en statistique et informatique décisionnelle » est inscrit au programme de la seconde année.

Au cours de la première année, les étudiants ont reçu des enseignements de mathématiques, probabilités et simulations, statistiques descriptives, étude statistique et enquête, une initiation à la statistique inférentielle, à l'ajustement de courbes et de séries chronologiques, programmation statistique. En informatique, ils ont étudié l'algorithmique et les bases de la programmation (en VB puis en Python), l'exploitation de données (disponibles au format tableur ou dans des bases de données relationnelles), les modèles et méthodes pour la structuration des bases de données, les technologies web (HTML, CSS et Javascript). Ils ont également déjà réalisé un projet tutoré par groupes de 3 à 5 étudiants, en général sur un jeu de données, qu'ils doivent exploiter sur un plan statistique.

En deuxième année, le module "étude de cas" se place au premier semestre (S3), où ont également lieu les cours suivants : analyse des données, estimation et tests, modèle linéaire en statistique, Système d'Information Décisionnel et développement d'applications décisionnelles en informatique. Il existe également un deuxième projet tutoré en deuxième année, qui au semestre 3 est consacré à l'appropriation du domaine d'application, la rédaction du cahier des charges, l'analyse comparative de solutions, la planification et la répartition des tâches, etc.

Le cours d'étude de cas en statistique et informatique décisionnelle est pensé comme un cours intermédiaire entre les cours traditionnels et le projet tutoré. Il est censé "mettre en oeuvre des techniques de statistique et d'informatique décisionnelle sur des données réelles". Il s'appuie en particulier également sur le cours de conduite de projet de première année. Il vise l'association "des compétences acquises en statistique et en informatique décisionnelle, sur la base d'un cahier des charges". Il doit être l'occasion d'approfondir l'utilisation de logiciels professionnels. Une contrainte, pour coller au plus près de la réalité terrain, est de partir de données réelles non qualifiées. Les séances doivent être encadrées par des enseignants et/ou intervenants professionnels des deux disciplines. Le travail donne lieu à une réalisation concrète, à une interprétation des résultats obtenus et à la rédaction d'un document de synthèse.

3 Format du cours

A Grenoble, le cours a été mis en oeuvre pour la première fois au premier semestre de l'année universitaire 2014-2015. Il a été mené par deux enseignantes (les deux auteurs de cette communication), l'une enseignante en statistique, l'autre en informatique. Il s'est déroulé sur 13 séances hebdomadaires de 2h, en salle machine. Trois séances sur les 13 ont pu se dérouler en présence des deux enseignantes : la première pour la présentation globale du cours et du projet support, la neuvième pour une évaluation intermédiaire sous forme de mini-soutenances, la dernière pour la finalisation de l'étude. Les autres séances ont été

menées soit par l'une, soit par l'autre selon les tâches à mener.

Le projet support a été présenté aux étudiants sous la forme d'un cahier des charges à réaliser pendant les séances de cours essentiellement, les travaux demandés devant être finis d'une séance à l'autre. Le cahier des charges suivait les recommandations données aux étudiants lors du cours de conduite de projet donné en première année. Il précisait le contexte et la définition du problème, les objectifs à atteindre (élaboration et interprétation de résultats, diffusion des résultats, les prestations attendues et livrables, les délais et la méthodologie proposée (pilotage du projet, répartition du travail), l'organisation et la planification, l'identification des risques et des critères d'évaluation interne. Les étudiants devaient constituer des binômes ou trinômes pour la durée du cours.

4 Contenu du cours

4.1 *Le cas support*

Le cas support pour ce cours est une enquête menée auprès de parents d'élèves de deux écoles de Grenoble. Le sujet de l'enquête est la mise en oeuvre de la réforme des rythmes scolaires et des activités périscolaires en 2013 à Grenoble. Les deux écoles participent à un même Projet Educatif Local (PEL), mis en oeuvre par une même MJC (Maison des Jeunes et de la Culture). L'enquête a été réalisée au printemps 2014 : un questionnaire papier a été distribués aux parents des deux écoles, puis les données ont été saisies par un groupe d'étudiantes de première année de STID dans le cadre de leur projet tutoré, à l'aide du logiciel Sphinx Online. Enfin, les données avaient déjà été analysées, mais de manière assez simpliste. Il nous a semblé que les données pouvaient être plus profondément exploitées, pourvu que les étudiants soient un peu plus cadrés.

Le questionnaire est assez complexe : composé de 8 pages, il débute par un rappel du dispositif existant (notamment pour fixer le vocabulaire utilisé), puis est composé de 29 questions, regroupées en plusieurs rubriques :

- Généralités (composition de la famille, classes et écoles des enfants)
- Activités du mercredi
- Activités périscolaires après la classe du lundi au vendredi
- Accompagnement scolaire (étude) et activités pédagogiques complémentaires (avec des enseignants)
- Récréative (activités gratuites dans l'école avec animateurs)
- Ateliers éducatifs proposés par la ville de Grenoble
- Activités proposées par la MJC
- Autres activités des enfants

Le questionnaire s'adresse aux familles. Certaines questions portent donc sur l'ensemble de la famille (exemples : nombre d'enfants scolarisés dans les deux écoles, situation de famille, Seriez-vous intéressé par un service de restauration le mercredi midi, etc), d'autres sur chacun des enfants de la famille (niveau scolaire, que fait votre enfant le mercredi à 11h30 ? si votre enfant est resté à la récréative, en est-il satisfait ? votre enfant a-t-il d'autres activités que celles citées ci-dessus (club sportif, conservatoire, ...) etc), d'autres enfin sur chacune des activités de chacun des enfants (nom de l'activité, organisme, jour de la semaine). Trois populations statistiques sont donc impliquées : celles des familles, celles des enfants, celles des activités. Enfin, les questions peuvent aussi bien être fermées, à choix multiples, qu'ouvertes.

Les données sont stockées dans un fichier Excel, lui-même exporté de Sphinx online. Une ligne du fichier correspond à un questionnaire papier, donc à une famille.

4.2 Travail demandé

Le premier travail demandé aux étudiants est un travail de correspondance entre les questions du questionnaire papier et les colonnes du fichier. Ils doivent être capables de retrouver dans quelles colonnes du fichier informatique se trouvent les réponses à telle question du questionnaire papier. Ils découvrent les incohérences de codage et sont amenés à les corriger (étape de nettoyage de données). Il leur était demandé ici de faire appel à leurs compétences dans l'utilisation du logiciel Excel et en programmation VB pour automatiser autant que possible la mise en évidence des incohérences et/ou leur correction.

Dans un second temps, les étudiants ont travaillé sur la restructuration du fichier de données pour formater correctement les trois populations statistiques en vue des analyses à mener. Il s'agissait d'obtenir 3 fichiers contenant respectivement :

- une ligne par famille
- une ligne par enfant (sans perte d'information sur sa famille)
- une ligne par activité d'enfant (sans perte d'information sur l'enfant et sa famille)

Pour cette étape, un travail de réflexion algorithmique devait d'abord être mené avant de proposer une implémentation de la solution sous forme de programme VB.

Ensuite a commencé la phase de traitement statistique proprement dite : chaque binôme ou trinôme d'étudiants s'est vu confié le traitement d'une partie du questionnaire. La première étape a été d'évaluer la qualité de l'enquête à travers le taux de retour des questionnaires en relation avec les effectifs dans chaque école. Dans le même ordre d'idée, les étudiants ont étudié la qualité de chaque réponse à une question à travers son taux de réponses (parmi les questionnaires rendus). Puis, les étudiants ont créé les tableaux et graphiques permettant de voir la répartition d'une variable réponse dans l'ensemble de la population, pour chaque groupe scolaire, pour chaque type d'école (maternelle/élémentaire), et enfin pour chacune des 4 écoles. Un commentaire et éventuellement un test statistique pour confirmer une différence entre deux écoles étaient attendus. Un rapport écrit sur ces résultats statistiques était demandé.

La dernière partie du travail consistait à concevoir et à mettre en oeuvre une solution pour la diffusion en ligne des résultats produits par le traitement statistique. La conception d'une base de données permettant de stocker tous les éléments du travail statistique (tableaux, graphiques, commentaires) était imposée. Les pages web présentant les résultats devait inclure des programmes de requêtage de la base de données. Une charte graphique, compatible avec celle du site de la MJC (en vue d'une intégration à son site web) était demandée.

Tout au long du module les étudiants ont été évalués à travers le rendu de travaux intermédiaires :

- Rendu d'un rapport sur les traitements statistiques
- Rendu d'un maquettage de l'application
- Développement d'une interface de consultation et rendu d'un rapport technique
- Passation d'un examen sur table

5 Attentes et résultats

Le cours a été conçu avec un certain nombre d'attentes pédagogiques. Nous les présentons

ainsi que le constat fait à l'issue de cette première expérience.

5.1 Mobiliser les connaissances et compétences acquises en première année

La capacité d'autonomie des étudiants sur des questions abordées en première année a été largement sur-estimée. Cela a été constaté tant en statistique qu'en informatique. En statistique, par exemple, certains réflexes en statistiques descriptives attendus des étudiants ne sont pas encore établis. L'intérêt d'étayer ses commentaires par un test statistique (étudié en parallèle dans un autre cours) n'est pas non plus ancré. Du point de vue informatique les difficultés majeures ont été constatées lors de l'élaboration des algorithmes de restructuration des fichiers, et lors de conception d'une base de données répondant aux besoins de l'application cible. Il semble clairement que le travail d'analyse d'une situation et de conception d'une solution soit un obstacle pour les étudiants à ce stade de leur formation. Le passage à la réalisation est de manière générale mieux abordé. Du point de vue informatique, nous veillerons à conserver les tâches d'ordre conceptuel car elles sont importantes pour la compréhension du schéma global dans lequel s'inscrivent les réalisations. Nous porterons toutefois une attention accrue à l'accompagnement des étudiants dans cette phase.

5.2 Prendre conscience de la réalité terrain

L'enquête choisie nous avait paru très riche en problèmes posés par des données réelles par rapport aux cas "scolaires" abordés habituellement en cours qui lissent une grande partie des difficultés. Elle s'est avérée souvent trop complexe, les problèmes de base étant déjà souvent très difficiles à résoudre pour nos étudiants. La confrontation à ces données a contribué à montrer aux étudiants une certaine réalité des conditions auxquelles pourront s'inscrire leurs futures missions professionnelles. Néanmoins, ce jeu de données en particulier s'est révélé inapproprié pour obtenir un bon équilibre dans lequel la réalité des données ne vient pas entraver l'atteinte des objectifs pédagogiques.

5.3 Contribuer à un projet global

Le projet dans son ensemble avait été présenté comme un objectif global à atteindre, chaque groupe d'étudiants devant y contribuer en traitant un aspect de l'enquête (une rubrique identifiée dans le questionnaire). Il devait donc y avoir des avancées en parallèle de chaque groupe avec un temps d'intégration des différentes productions. Par exemple, le projet initial ambitionnait que les rapports sur les traitements statistiques produits par chaque groupe suivent un modèle de document fournis (avec numérotation automatique des titres par exemple) afin que la fusion des fichiers soit facilitée. De même, les modèles conceptuels de base de données conçus par chaque groupe devaient ensuite donner lieu à un temps de discussion pour s'accorder sur un modèle commun. Chaque groupe implémenterait ce modèle puis le peuplerait avec ses données (et l'utiliserait pour la partie web) avant de fournir un script .sql d'export qui aurait servi à alimenter la base commune.

Globalement, les étudiants comprennent l'importance de mener à bien, correctement, dans le respect des spécifications et dans les temps, une tâche confiée dans l'intérêt du projet principal. Néanmoins, on relève un certain manque de rigueur dans la réalisation qui a pénalisé parfois la réussite de l'objectif global (ex. la création d'un rapport intégrant les productions des groupes). Par ailleurs, du fait des difficultés rencontrées par les étudiants (cf. 5.1), certains aspects ont été abandonnés en cours de module faute de temps (ex. l'importation des données dans une seule base qu'auraient interrogée les différentes pages web des groupes)

5.4 S'impliquer pour la réussite du projet

Lors des toutes premières séances, nous avons constaté un intérêt des étudiants pour ce projet. Ils semblaient motivés par l'objectif et volontaires. Peu à peu, sans que cela soit

cependant très net, ils ont paru moins mobilisés. Deux facteurs nous semblent expliquer cela. Parallèlement aux difficultés rencontrées qui ont pu être source de découragement chez eux, le fait d'avoir à travailler selon un rythme hebdomadaire a été préjudiciable. Avec le recul, il nous paraît bien plus efficace de travailler de manière concentrée, sur une semaine consacrée à cette étude de cas, en mode projet.

5 Conclusion

Ce cours nous a paru une excellente occasion de remettre en oeuvre les techniques de statistique et d'informatique vues en première année à travers un cas concret. Les préconisations du PPN pour ce cours ont bien été respectées. La présentation sous forme de projet a permis d'impliquer les étudiants et de différencier ce cours des cours traditionnels. Cependant, le cas choisi s'est avéré souvent trop compliqué alors que des compétences élémentaires pour des deuxième année n'étaient pas encore acquises. Nous envisageons donc de repartir sur une enquête plus simple l'an prochain, tout en gardant l'esprit du déroulement de ce cours que nous organiserons sur une semaine dédiée.

Bibliographie

[1] Ministère de l'enseignement supérieur et de la recherche, (2013), Diplôme Universitaire de Technologie STATISTIQUE ET INFORMATIQUE DECISIONNELLE *Programme Pédagogique National*, http://cache.media.enseignementsup-recherche.gouv.fr/file/24/23/5/PPN_STID_255235.pdf.