NOTES DE STATISTIQUE ET D'INFORMATIQUE

2012/3

LA COLLECTE DE L'INFORMATION
POUR QUI ? POURQUOI ? COMMENT ?

J. J. CLAUSTRIAUX

Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
GEMBLOUX
(Belgique)

LA COLLECTE DE L'INFORMATION. POUR QUI ? POURQUOI ? COMMENT ?

J. J. Claustriaux *

RÉSUMÉ

Cette note tente de répondre succinctement aux trois questions : pour qui, pourquoi et comment collecter l'information en vue de mettre au point un système d'aide à la décision.

SUMMARY

This paper tries briefly to answer three questions : who, why and how to collect information to develop a system for decision support.

1. INTRODUCTION

Cette publication reprend le corps d'une conférence présentée le 21 novembre 2012 dans le cadre d'une journée d'études intitulée « La collecte de l'information et le développement des systèmes d'aide à la décision, un appui à une agriculture en constante évolution », organisée par le Centre wallon de Recherches agronomiques [1].

Après cette introduction (paragraphe 1), la notion d'information est décrite de façon générale (paragraphe 2).

Ce terme étant mieux circonscrit, il est alors possible de tenter de répondre aux trois questions (paragraphes 3, 4 et 5) concernant la collecte de l'information : pour qui, pourquoi, comment ?

Enfin, une conclusion est proposée, elle aussi sous la forme d'une question (paragraphe 6).

*Professeur ordinaire à l'Université de Liège, Gembloux Agro-Bio Tech.
1. Rue de Liroux, 9, B-5030 Gembloux

2. L'INFORMATION ?

1° L'information est un terme qui vient du verbe latin *informare*; il signifie « donner forme à » ou « se former une idée de ».

C'est un concept ayant de nombreux sens, dont certains sont très complexes comme « Robert est un homme ».

Pour notre part, l'information se confond avec la notion de donnée. Celle-ci est numérique ou alphanumérique, c'est-à-dire qu'elle peut être utilisée pour effectuer des opérations arithmétiques et logiques. Une unité de grandeur peut lui être associée.

Pour la mise au point de phénomènes complexes, on fait aussi référence à des métadonnées. Une métadonnée est une donnée de données, une donnée à propos de données ou une donnée qui décrit ou définit une autre donnée. Un exemple simple est la métadonnée qui réunit le rendement d'une parcelle de froment associé à des photographies de la parcelle récoltée, des relevés météorologiques, une description de l'environnement cultural, etc.

2° Toute donnée est le résultat délibéré d'une observation ou d'une simulation. C'est la donnée qui constitue la matière première pour la mise au point de modèles, en particulier d'un système d'aide à la décision.

Enfin, l'observation est réalisée pour autant qu'un système pouvant exprimer une donnée soit disponible comme par exemple une plante, un animal, un territoire, etc. Il est appelé unité d'observation, unité d'échantillonnage si seulement quelques individus d'une population peuvent faire l'objet d'observations ou unité expérimentale dans le cas d'une expérimentation.

A ce propos, il convient d'insister sur un élément fondamental. Souvent, le chercheur regarde avec attention les données, sur lesquelles il applique éventuellement des méthodes statistiques sophistiquées, en particulier celles qui sont proposées par les logiciels. Rarement, il consacre un temps suffisant pour réfléchir préalablement à son domaine ou à son « espace » temps d'observation et, en particulier, à la qualité des unités à observer. Pourtant, ce sont certaines d'entre elles qui devront constituer l'échantillon, par exemple aléatoire et simple, pour exprimer des données qui contribueront à la mise au point d'un système d'aide à la décision utile.

Ce n'est que si les données ont été collectées selon une procédure adéquate que l'analyse statistique pourra débuter.

Pour étayer le propos, prenons comme exemple l'article de Seralini *et al.* [2012] dont les conclusions au sujet de l'usage d'un herbicide très connu ont dernièrement défrayé les chroniques .

Au niveau du traitement statistique des données, la description des méthodes est assez détaillée :

- analyse en composantes principales (ACP),
- régression PLS (PLS),
- régression orthogonale PLS (OPLS),

– analyse discriminante associée à la régression orthogonale PLS (OPLS - DA).

Par ailleurs, les facteurs de l'expérimentation sont aussi décrits, dont les dix objets et les deux sexes, ainsi que les unités expérimentales (100 rats par sexe répartis dans des cages de deux rats). Mais absolument rien n'est expliqué au sujet de l'allocation des objets aux unités, élément pourtant essentiel pour déterminer le modèle qui permettra d'évaluer la variabilité expérimentale, c'est-à-dire, notamment, de comprendre comment l'expérimentation a été menée pour contrôler l'hétérogénéité éventuelle du domaine expérimental et surtout comment les objets ont été répartis dans le domaine expérimental pour assurer une reproductibilité suffisante des résultats de l'expérience.

3° Poursuivons cette évocation sur l'information sans nous attarder sur les types de données : comptages, mensurations, attributs, données alternatives, données ordinales, données nominales, etc.

Néanmoins, il faut savoir qu'une réflexion *a priori* sur ce sujet est aussi essentielle car n'y a-t-il pas un lien étroit entre les types de données, les distributions de probabilité, le nombre de données à observer et, par conséquent, le nombre de répétitions à considérer [Dagnelie, 2012], etc. ?

Rares sont les ouvrages qui consacrent quelques paragraphes sur l'importance de la prise en compte de cet aspect pour réfléchir aux modalités d'organisation optimale de la collecte de l'information. Dagnelie [2007] est sans doute un des très rares auteurs originaux à ce propos.

Par ailleurs, reconnaissons que des outils existent pour la détermination du nombre d'informations à collecter (formules, tables, abaques, logiciels). Mais, ils sont rarement utilisés en pratique peut-être parce qu'ils sont trop abstraits, trop difficiles à mettre en oeuvre en raison des connaissances *a priori* exigées ou que les propositions qu'ils suggèrent s'inscrivent totalement en dehors des moyens mis à disposition ou du temps disponible. Néanmoins, est-ce scientifiquement une raison suffisante pour ne pas se référer à la théorie, quitte à justifier pourquoi elle n'est pas suivie ? Cela aurait au moins le mérite de comprendre les limites de l'usage judicieux d'un système d'aide à la décision.

C'est certainement aussi en raison du nombre considérable de cas particuliers à résoudre pour définir un mode de collecte de l'information qu'il y a cet écart entre théorie et pratique.

Pour illustrer cette diversité de situations, résumons deux exemples pourtant très proches. Ainsi, si une première méthode d'échantillonnage d'un sol pour une analyse des reliquats d'azote prévoit notamment de partager la parcelle en trois zones homogènes et, dans chaque zone, de faire six à huit carottages par horizon [Fontaine, 1987], une seconde méthode d'échantillonnage d'un sol sur des grandes parcelles en vue d'une analyse physico-chimique suggère de déterminer une zone homogène, de fixer un point facilement repérable dont les coordonnées seront relevées et d'y piqueter les emplacements de douze prélèvements sur la circonférence d'un cercle de 10 à 15 mètres autour du point [Taureau, 1987].

Notons cependant que la démarche dite de qualité a introduit dans les

processus de productions répétitifs ou d'analyses à la chaîne, des normes pour le contrôle statistique de la qualité qui souvent imposent les modalités précises relatives à la collecte des données.

4° Enfin, l'important ce n'est pas seulement de collecter de l'information. Faut-il encore que celle-ci soit obtenue sous une forme qui permette ultérieurement le traitement des données, y compris la vérification des données et des conditions d'application des méthodes employées. Le codage des données est donc aussi important à considérer *a priori* [Pearce, 1983].

3. LA COLLECTE DE L'INFORMATION : POUR QUI ?

1° La réponse à la question « la collecte de l'information : pour qui ? » semble évidente, à savoir pour le chercheur, au sens générique du terme.

Certes, la mise au point d'un modèle ne nécessite pas obligatoirement une collecte d'informations. Cependant, il ne me semble pas possible de concevoir un système d'information sans avoir préalablement observé le phénomène qu'on souhaite modéliser pour identifier les variables influentes et leur limite de variation.

Par ailleurs, même si le modèle n'est construit que sur des bases théoriques, il devra être validé et à ce moment, le chercheur devra disposer de données qu'il ira collecter.

L'agronomie n'est pas la physique ; ainsi, le modèle unique d'évolution des rendements en fonction des fumures n'existe pas, sauf sous sa forme générale inutilisable en pratique.

Le chercheur a donc besoin de collecter de l'information si son objectif quitte le cadre strictement théorique, peu importe le degré de sophistication de son système d'information et ce d'autant plus que ce dernier est complexe.

2° Collecter de l'information, ce n'est certainement pas pour le système d'information en tant que tel, car celui-ci n'aura jamais raison, il ne sera jamais exact, il ne se trompera jamais ; aucun système d'information n'est faux !

Par contre, celui qui conçoit le système peut se tromper ; il est entièrement responsable du choix de son système et surtout des hypothèses qui le supportent [Legay, 1997].

Le concepteur du système se trompera d'autant moins que la collecte de l'information sera réfléchie et judicieuse, en particulier que les caractéristiques des unités d'échantillonnage seront le mieux décrites.

Finalement la question n'aurait-elle pas dû être : la collecte de l'information : sur qui ?

4. LA COLLECTE DE L'INFORMATION : POURQUOI ?

A la question « la collecte de l'information : pourquoi ? », la réponse semble aussi évidente.

1° De façon générale, l'organisation de la collecte de l'information de nature biologique, comme c'est le cas dans la recherche agronomique, est un élément crucial pour la construction d'un système d'aide à la décision représentatif tenant compte de trois éléments : la variabilité des organismes vivants, la variabilité des réactions de ces organismes aux effets imposés ou aléatoires et l'hétérogénéité du milieu dans lequel ils croissent et ils se développent.

2° Par ailleurs, cette collecte de l'information est indispensable et continue car le domaine agronomique, l'agriculture en particulier, ne cesse d'évoluer suite aux changements permanents qu'ils soient sociétaux, économiques, techniques, environnementaux. Constamment, elle pose de nouvelles questions à la recherche qui, même en matière de collecte des données, doit innover.

A titre d'exemple, la mise à disposition de « drones », ou aéronefs ultralégers, commandés à distance, permettant aujourd'hui à certains chercheurs d'observer dans les trois dimensions un champ, une forêt, une savane, va influencer les modes de collecte de l'information et bien entendu tout le traitement des données par l'introduction plus importante encore de la spatialité et du caractère plus ponctuel ou précis de cette information.

Cependant, une question doit être posée : est-ce que collecter une quantité aussi monstrueuse d'informations est nécessaire pour la mise au point d'un système d'aide à la décision adapté aux conditions de la pratique, qui selon l'adage doit finalement trouver son arbre dans la forêt ?

3° Ajoutons encore que la collecte de l'information est devenue un véritable enjeu économique dans certains secteurs, certes proches de la publicité, du marketing et du commerce. C'est devenu le métier de certaines grandes entreprises qui cherchent à comprendre le comportement du consommateur pour ensuite le piéger en le poussant à acheter ou à consommer. La recherche de niches paye actuellement : faites un clic avec votre souris et vous serez fiché sans le savoir.

L'agriculture ne fait pas exception. Ainsi pour des raisons économiques, ne cherche-t-on pas à déterminer quelles sont les catégories de consommateurs de produits alimentaires intéressés par les circuits courts, les produits biologiques, etc. ?

5. LA COLLECTE DE L'INFORMATION : COMMENT ?

1° Comment collecter l'information est une question pour laquelle la plupart des réponses se trouvent depuis longtemps dans la littérature pour ce qui concernent les techniques d'échantillonnage [Yates, 1951], même si certains ouvrages sont plus récents et n'apportent pas d'innovation particulière [Scheaffer *et al*, 2012].

Outre l'échantillonnage complétement aléatoire déjà évoqué et son cas particulier l'échantillonnage systématique, citons l'échantillonnage stratifié, l'échantillonnage à plusieurs degrés, l'échantillonnage à probabilités proportionnelles à la taille des unités, l'échantillonnage par quota, etc.

Il en est de même en expérimentation ; de nombreux dispositifs complets ou incomplets, partiels ou fractionnaires, peuvent être implantés en fonction du type d'expérimentation (essais de variétés, essais phytotechniques, essais pérennes, essais de pâturage), des spécificités des unités expérimentales, de l'hétérogénéité expérimentale, des moyens et du temps disponibles, etc. [Dagnelie, 2012].

Cependant, l'évolution de la recherche va compliquer le travail du statisticien qui jusqu'à présent dialoguait souvent seul face au chercheur venu le consulter pour résoudre un problème bien circonscrit d'échantillonnage ou d'expérimentation.

En effet, la recherche agronomique semble s'orienter vers l'étude de système plus complexe comme par exemple l'étude des cultures associées ou la modélisation des systèmes de productions agricoles pour pouvoir apporter un éclairage objectif aux conséquences des changements climatiques annoncés.

Dans ce cadre, quelle méthode de collecte de l'information mettre en place ? Je n'ai aucune réponse *a priori*. Toutefois, je pense que davantage qu'auparavant, ce sera toute une équipe de chercheurs, parmi lesquels un statisticien au moins, qui devra composer une ou des réponses, en y incluant davantage le contexte économique et l'aspect démonstratif [Petersen, 1994].

2° Il me semble aussi qu'une partie de la réponse concerne l'attention qu'il convient d'apporter aux instruments utilisés pour collecter l'information.

Les outils informatiques et les capteurs de données en temps réels et à distance ont fait largement leur entrée dans le champ de la recherche agronomique. On ne peut que s'en réjouir surtout si de plus un robot peut se substituer au chercheur pour collecter l'information.

Cependant, ces appareils sophistiqués à usage limité et qui nécessitent souvent un temps de conception non négligeable, accumulent des quantités phénoménales de données souvent dans un nombre limité de situations, si ce n'est pas en un seul point d'observation. Dès lors, comme déjà signalé, qu'en est-il de la reproductibilité des mesures et en conséquence de la qualité des conclusions ?

En pratiquant ainsi, ne fait-on pas fi de ce que nous apprend tout simplement par exemple la variance de la moyenne dans le cas d'un échantillonnage à plusieurs degrés, à savoir qu'il convient de réduire autant que possible le nombre d'unités du degré supérieur, c'est-à-dire de limiter le nombre d'observations en un point ou répétabilité des mesures et, en conséquence, d'augmenter les nombres de points ?

6. EN GUISE DE CONCLUSION

Finalement, faut-il conclure ?

Il n'est pas possible de conclure sur le sujet très vaste de la collecte de l'information si ce n'est pour rappeler une fois encore que la moitié du temps consacré à la mise au point d'un système d'information devrait être dédiée aux réflexions préalables sur la collecte de l'information, notamment sur le choix des unités à observer, point de départ de la définition des objectifs d'une recherche.

BIBLIOGRAPHIE

Dagnelie P. [2007]. *Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique.* 2e édition. Bruxelles, De Boeck, 511 p.

Dagnelie P. [2012]. *Principes d'expérimentation.* Gembloux, Presses agronomiques, 413 p.

Fontaine. [1987]. Méthode d'échantillonnage d'un sol pour une analyse des reliquats d'azote. *In* Gouet H., Beaux M-F, catalogue de fiches « méthodes ». Arvalis, Institut du Végétal, Milieu 01068.

Legay J.M. [1997]. *L'expérience et le modèle.* Paris, INRA, 110 p.

Pearce S.C. [1983]. *The agricultural field experiment. A statistical examination of theory and practice.* New York, Wiley, 335 p.

Petersen R.G. [1994]. *Agricultural field experiments.* New York, Dekker, 409 p.

Seralini G.-E., Clair E., Mesnage R., Gress S., Defarge N., Malatesta M., Hennequin D., Spiroux de Vendômois J. [2012]. Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize. *Food Chem. Toxicol*, 50, 4221-4231.

Scheaffer R. L., Mendenhall W., Lyman Ott R., Gerow K. G. [2012]. *Survey Sampling.* 7th ed. Boston, Brooks/Cole, 436 p.

Taureau J.C. [1987]. Méthode d'échantillonnage d'un sol sur des grandes parcelles pour une analyse physico-chimique. *In* Gouet H., Beaux M-F, catalogue de fiches « méthodes ». Arvalis, Institut du Végétal, Milieu 01073.

Yates F. [1951]. *Méthodes de sondage pour recensements et enquêtes.* Paris, Dunod, 335p.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant de l'Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie de l'Université de Liège – Gembloux Agro-Bio Tech et de l'Unité Systèmes agraires, Territoire et Technologies de l'Information du Centre wallon de Recherches agronomiques (Gembloux - Belgique).

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
Avenue de la Faculté d'Agronomie, 8
B-5030 GEMBLOUX (Belgique)
E-mail : sima.gembloux@ulg.ac.be

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

http://www.gembloux.ulg.ac.be/si/

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

CHARLES C. [2011]. Introduction aux ondelettes. *Notes Stat. Inform*. (Gembloux) 2011/1, 22 p.

CHARLES C. [2011]. Introduction aux applications des ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/2, 35 p.

PALM R., BROSTAUX Y. et CLAUSTRIAUX J. J. [2011]. Macros Minitab pour le choix d'une transformation pour la normalisation de variables. *Notes Stat. Inform.* (Gembloux) 2011/3, 15 p.

PALM R., BROSTAUX Y. [2011]. La régression logistique avec Minitab. *Notes Stat. Inform*. (Gembloux) 2011/4, 15 p.

PALM R., BROSTAUX Y., CLAUSTRIAUX J. J. [2011]. Inférence statistique et critères de qualité de l'ajustement en régression logistique binaire. *Notes Stat. Inform.* (Gembloux) 2011/5, 32 p.

CLAUSTRIAUX J. J., PALM R., FERRANDIS-VALLTERRA S., BROSTAUX Y. et PLANCHON V. [2012]. Tables de contingence à trois dimensions : aspects théoriques, applications et analogie avec l'analyse de la variance à trois critères de classification. *Notes Stat. Inform*. (Gembloux) 2012/1, 19 p.

PALM R., CHARLES C., CLAUSTRIAUX J. J. [2012]. La représentation d'une matrice par biplot. *Notes Stat. Inform*. (Gembloux) 2012/2, 22 p.

D/2012/2371/3