Ressources SPSS pour l'exploration statistique de données

Exploration statistique des données par le logiciel SPSS

Introduction :

I) Initiation au logiciel
- présentation du logiciel SPSS
- découverte de SPSS, manipulation de données : compute, recode
- traitement des sorties : Tableaux , Graphiques

II) Premières analyses : statistique descriptive
- l'histogramme
- la boîte à moustaches
- Présentations et résumés

III) Les tests d'hypothèses statistiques
- Etude d'un échantillon : one sample
- comparaison de deux moyennes
- Analyse de la variance : anova – F test
- tableaux croisés : tests du Chi2

IV) Les méthodes statistiques explicatives : essais de modélisation
- régression simple
- régression multiple
- régression logistique

V) Ecarts aux hypothèses du modèle linéaire :
- asymétrie,
- points aberrants,…
- données non normales (GLM)

VI) L'exploration multivariée
- ACP
- Classification
- Analyse discriminante
- AFC

VII) Les séries temporelles

VIII) Les développements
- Les données géographiques et la commande Maps
- Le Data mining et le texte mining

IX) Conclusions

Bibliographie

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc

Youssfi El Kettani - Juillet 2008

youssfielkettani@univ-ibntofail.ac.ma

Ressources SPSS pour l'exploration statistique de données

Introduction

Etant donné la diversité des notions abordées en traitement statistique des données, le présent document est une synthèse qui a pour principal objet de faciliter l'orientation et la progression du lecteur à travers les sites web indiqués tout au long du document et à travers les références bibliographiques rassemblées à la dernière page et auxquelles le document renvoie en indiquant entre [ ] le numéro de la référence. Il est fortement conseillé de pratiquer les manipulations du logiciel, appliquant pour chaque méthode, les notions abordées sur les exemples pratiques qui accompagnent les documents ou à défaut ceux intégrés au logiciel (études de cas et jeux de données).

Ainsi chacun pourra prendre en charge sa propre exploitation des documents et exemples en fonction d'objectifs de révisions ou d'approfondissements qu'il se fixe.

Enfin, étant donné le dynamisme que connaît le sujet et son évolution permanente, il est naturellement très intéressant de mettre à jour, voire compléter les documents bibliographiques ci-joints ainsi que les sites web indiqués, en effectuant de temps à autre des recherches de nouveaux documents, notamment sur la toile du web où les sites de plusieurs professeurs et laboratoires de recherche offrent une multitude de ressources pédagogiques.

I) Initiation au logiciel SPSS

A) présentation du logiciel SPSS

Il est intéressant de lire à titre d'introduction générale:

- l'article sur SPSS fait dans l'encyclopédie sur l'Internet, Wikipedia : http://fr.wikipedia.org/wiki/SPSS

- ainsi que la présentation faite sur le site de SPSS Maghreb http://www.spssmaroc.ma/spss/data_analysis.php

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc

Youssfi El Kettani - Juillet 2008
youssfielkettani@univ-ibntofail.ac.ma

Par ailleurs, les possibilités de SPSS sont gigantesques et le meilleur moyen pour bien cerner chacune des possibilités qu'offre le logiciel est de pouvoir se référer très fréquemment au manuel de référence [1]. Ce document est en effet, complet avec de nombreuses illustrations par des boîtes de dialogue aidant à la compréhension des manipulations à faire pour chaque commande.

Toutefois une bonne exploitation du logiciel SPSS passe nécessairement par un minimum de connaissances exactes des méthodes statistiques. A cet effet, il est proposé tout le long du présent document des liens hypertextes vers des documents de cours et/ou d'exercices se rapportant aux différentes parties traitées.

Un autre recours qu'il faut signaler ici est la fonction « Aide » intégrée au logiciel. On y trouve non seulement un index de recherche accompagné d'un lexique très développé pour chaque rubrique, mais aussi plusieurs exemples et cas d'étude dont on peut suivre les démonstrations pas à pas illustrant ainsi la grande partie des principales commandes et routines du logiciel.
Il est utile de consulter à ce sujet le chapitre 2 de [1].

B) Découverte de SPSS, manipulation de données :
Dès les premiers contacts avec SPSS on s'aperçoit que nous allons devoir nous familiariser avec un certain nombre de fonctionnalités du logiciel, citons notamment :

- Les différentes interfaces de SPSS
    - L'éditeur de données
        - Le mode « variables »
        - Le mode « données »
        - Les étiquettes
        - Les données manquantes
    - La barre d'outils ; les commandes
        - Les boîtes de dialogue
        - L'éditeur de syntaxe
    - L'éditeur de résultats
- L'introduction et les transformations de données

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc
Youssfi El Kettani - Juillet 2008
youssfielkettani@univ-ibntofail.ac.ma

- Les différents types de variables
- La saisie des données et des caractéristiques des variables
- La sélection d'individus selon une condition : if
- La transformation des variables : la commande compute

Tous ces points peuvent être repris avec beaucoup d'intérêt dans le document de Donald Long intitulé une introduction à SPSS, en format pdf et qui est disponible à l'adresse : http://www.umoncton.ca/longd03/
Ou encore dans « An Introduction to *SPSS* for *Geographers* » qu'on peut télécharger de www.geog.umontreal.ca/donnees/geo1512/**SPSS**%20Handbook%202004.doc ainsi que le document d'exercices du même auteur, Dr. Stewart Barr, disponible dans: www.projects.ex.ac.uk/ebrg/Exercise%20handout.doc
Ou bien dans le document de SPSS Inc, cf bibliographie [2]:
Il y a également un document récent qui traite par ailleurs des possibilités de programmation dans SPSS est Spss For Dummies, [3].

II) Premières analyses

Présentations et résumés

La statistique descriptive touche tous les aspects de description de présentation et de résumés de l'information contenue dans un ou plusieurs échantillons.
La pratique de la statistique descriptive par SPSS est présentée dans le document en format PDF à l'adresse suivante : http://www.pifo.uvsq.fr/pedagogie/bime/spss_guide.pdf
Et on peut trouver beaucoup de détails dans [1] et aussi dans [14], ouvrage particulièrement conçu pour s'initier aux traitements statistiques des données.

Enfin il est très intéressant pour la pratique sur SPSS de consulter les chapitres 14 et 15 du livre de référence : [1]

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc

Youssfi El Kettani - Juillet 2008
youssfielkettani@univ-ibntofail.ac.ma

Ou encore de se référer aux documents très complets cités dans [4] et [5]de la bibliographie.

La statistique descriptive permet, à l'aide de tableaux et graphiques, de visualiser les variables étudiées, d'abord une par une puis certains tableaux et graphiques permettent de faire l'étude simultanée de deux variables.
Par ailleurs des indicateurs numériques comme la moyenne, le mode, l'étendue, l'écart-type ou le coefficient de corrélation synthétisent au maximum l'information contenue dans les variables étudiées.
Par ailleurs, un certain nombre de graphiques sont très utiles dans la description des variables et de la manière dont ils sont répartis. Citons plus particulièrement les histogrammes et les diagrammes de Tuckey ou boîtes à moustaches

Les histogrammes

Dans le cas d'une variable continue, on peut construire un histogramme des effectifs. Si les classes sont de même amplitude, en plaçant en ordonnée les effectifs on obtient des rectangles dont la surface est proportionnelle à l'effectif associé.
Le cours st@atnet sur Internet présente dans son chapitre sur la statistique descriptive la notion d'histogramme de façon détaillée. On peut consulter ce cours à l'adresse :

http://www.agro-montpellier.fr/cnam-lr/statnet/cours.htm

Les boîtes à moustaches :

Un document complet sur cette question est de Leguen Dominique, publié en 2001 et qu'on peut également télécharger du site dédié aux documents pour la statistique:
www.wikistat.ca .

III) Les tests d'hypothèses statistiques

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc

Youssfi El Kettani - Juillet 2008
youssfielkettani@univ-ibntofail.ac.ma

La notion de tests d'hypothèses statistiques est primordiale lorsqu'on veut intégrer les données statistiques dans le processus de prise de décision associée à un calcul de risque d'erreur.
le cours st@atnet dejà cité, http://www.agro-montpellier.fr/cnam-lr/statnet/cours.htm est tout indiqué pour s'initier aux concepts de base des tests d'hypothèses.
Par ailleurs, un document intitulé « aide mémoire » peut être téléchargé à l'adresse :
http://www.unifr.ch/ipg/cours/SemLic/manuelSPSS.doc.pdf Il présente l'utilisation des principaux tests par SPSS. Et un exemple d'utilisation est sommairement présenté dans le document intitulé " testing hypotheses using SPSS.pdf." disponible à l'adresse :
http://people.richland.edu/james/lecture/spss/testing/testing.pdf
Et pour un document complet qui traite par SPSS les tests liés à des modèles statistiques comme l'analyse de la variance, la régression multiple ou l'analyse multivariée, on peut se référer au livre [6]
Les tests non paramétriques sont traités dans le document intitulé estimation non paramétrique par SPSS de Jean-Marie Le Goff1 et Yannic Forney1 Décembre 2003 qu'on peut obtenir à l'adresse :
http://www2.unil.ch/pavie/documentation/methodesnonparametriques.pdf

IV) Les méthodes statistiques explicatives : essais de modélisation

- La régression simple

Un exposé de la statistique par SPSS avec un rappel de certains tests d'hypothèses suivi d'un développement de la régression simple par SPSS est proposé par Marjorie Noël, Stéphanie Parent et Catherine Shedleur de HEC :
http://zonecours.hec.ca/documents/200597.techniquesstatistiquespresentation.ppt

Par ailleurs il est très intéressant de suivre et de bien comprendre le traitement du cas concret de régression simple présenté par Michel Tennenhaus de l'école HEC ; dans

https://studies2.hec.fr/jahia/webdav/site/hec/shared/sites/tenenhaus/acces_anonyme/home/fichier_ppt/r%C3%A9gression_simple%20hec.ppt
Et il ne sera pas inutile de refaire soi même les traitements de cet exemple à partir des données reproduites en annexe1.

- La régression multiple

C'est la généralisation naturelle de la régression simple au cas où on veut expliquer une variable réponse à partir de plusieurs variables explicatives.
A ce niveau on peut lire des documents plus développés sur la régression dans [7] Ou encore dans le document de SPSS Inc. [8].

- La régression logistique

Quand la variable réponse ne prend qu'un petit nombre de valeurs possibles : 2 ou 3 et que les variables explicatives sont continues et donnent lieu naturellement à des prédictions continues, il est indispensable de procéder à la transformation dite logit pour adapter la variable réponse.

Le document suivant de Mélanie Bourdon, Ève-Marie Filiatrault et Évelyne Robineau fait un développement très riche de la méthode :
http://zonecours.hec.ca/documents/200594.Regressionlogistique(versionfinale).ppt

V) Ecarts aux hypothèses du modèle linéaire

Nous rappelons ici les trois situations d'écarts aux hypothèses du modèle linéaire rencontrées lors du séminaire :

- l'asymétrie,
- les points aberrants,…
- et les données non normales

Pour l'asymétrie, le diagnostic de cette situation se fait à partir du coefficient d'asymétrie (skewness en statistique

descriptive) ainsi que par le diagramme de Tuckey ou encore par l'histogramme
Le traitement de données à forte asymétrie se fait par transformation de variables comme il est indiqué sur l'exemple Transformer dans le cours en ligne sur le modèle linéaire de Marc Bourdeau qu'on peut télécharger à l'adresse suivante : http://www.agro-montpellier.fr/cnam-lr/statnet/cours_autre.htm .

Pour les points aberrants, il s'agit ici tout simplement de rappeler l'importance de l'examen minutieux des données comme on peut le voir sur l'exemple de régression intitulé « concentration » dans le cours en ligne sur le modèle linéaire, précité de Marc Bourdeau. http://www.agro-montpellier.fr/cnam-lr/statnet/cours_autre.htm

Et pour les données non normales, comme par exemple les données catégorielles, les données de comptage ou les données binaires, il y a lieu suivant la situation faire de la modélisation non linéaire, par exemple les modèles linéaires généralisés pour les données de comptage ou la régression logistique pour les données binaires. Pour le traitement par SPSS de la régression logistique que nous avons vue en IV ou encore des modèles linéaires généralisées, il est utile de consulter [8].

De façon plus générale on peut voir pour des données catégorielles [9] ou bien [10] qui contiennent diverses applications traitées par SPSS

VI) Les explorations multivariées

Un survol général des méthodes multivariées est présenté par John Zhang dans
http://ece.ut.ac.ir/classpages/F83/IPS/stat/spss/Multivariate%20Data%20Analysis%20Using%20SPSS.ppt

Ou encore le document du Professeur Besse disponible sur son site : http://www.math.univ-toulouse.fr/~besse/enseignement.html

- L'Analyse en composantes principales

C'est la méthode de base en analyse des données multivariées. Elle consiste à définir un ou deux plans principaux sur lesquels le nuage de points, souvent volumineux et appartenant à un espace mathématique de grande dimension, peut être projeté avec une perte acceptable de l'information contenue dans le nuage. Ces projections sur les plans principaux donneront des représentations interprétables et exploitables de la configuration du nuage.
Pour l'analyse en composantes principales par SPSS l'article de Dominique Desbois illustre très bien les diverses notions: http://www-rocq.inria.fr/axis/modulad/archives/numero-20/Desbois/uneintroduction.pdf
Il est très conseillé de reprendre individuellement cette étude de cas très pédagogique prise de [11]. A cet effet les données sont reproduites en annexe2.

Rappelons ici que la rubrique d'aide intégrée à SPSS est souvent très utile pour bien comprendre certaines procédure. Dans le cas de l'ACP, l'aide propose d'étudier le fichier de données car_sales.sav inclus dans le répertoire Programme Files > SPSS > tutorial > Sample files. Ce répertoire contient par ailleurs plusieurs exemples et études de cas traités par la rubrique d'aide. Nous proposons ici de reprendre individuellement le traitement de cet exemple par spss.

- les classifications automatiques

Ce sont les méthodes indiquées pour définir des classes de ressemblance dans une population. Elles sont très utilisées en markéting, notamment pour segmenter un marché. Un

document qui introduit les méthodes de classification et leur utilisation par SPSS est intitulé « création de Typoleogies sous SPSS » de Lemoal. Il peut être téléchargé à l'adresse : http://www.lemoal.org/download/spss/Typologies.pdf
Une étude de cas en classification est intitulée Epistémologie et Méthodologie Quantitative , SPSS Projet Pommes. Elle peut être téléchargée à l'adresse : https://studies2.hec.fr/jahia/webdav/site/hec/shared/sites/tenenhaus/acces_anonyme/home/projet/projet_pomme_soukeya.pdf
Enfin le site du professeur Gey propose un certain nombre de ressources pédagogiques sur la classification : http://www.math-info.univ-paris5.fr/%7Egey/ens.html

- L'analyse discriminante

Cette méthode d'exploration multivariée est également considérée comme une méthode explicative. Elle consiste à déterminer la combinaison linéaire de variables X1,…,Xk qui soit à même de départager une population . On peut effectuer une analyse discriminante pour confirmer une classification automatique. On obtient ainsi par l'analyse discriminante un modèle de prédiction qui permet d'affecter chaque nouvel individu à une classe. Un article qui introduit l'analyse discriminante et son utilisation par SPSS est présenté par Lemoal dans http://www.lemoal.org/download/spss/Analyse%20Discriminante.pdf
Nous proposons également de consulter l'article sur ce sujet intitulé « Analyse discriminante » de Christine Decaestecker & Marco Saerens : http://www.isys.ucl.ac.be/etudes/cours/linf2275/07cours1.pdf
ainsi que le chapitre de Tufféry sur les méthodes prédictives : http://data.mining.free.fr/cours/Predictives.PDF , pages 82 et suivantes.

- L'analyse factorielle des correspondances

C'est l'application de l'Analyse en composantes principales à des données de comptages présentées sous forme de tableau croisé. Cette application se fait par le choix d'une métrique convenable qui sera à la base des calculs des coordonnées de chaque point du nuage et de la détermination des facteurs et plans principaux. Un document qui présente parfaitement le sujet peut être consulté à l'adresse suivante:
http://www.mapageweb.umontreal.ca/durandc/Enseignement/MethodesQuantitatives/corresp1.htm
De plus il établit des liens utiles.

Par ailleurs, à l'instar de l'ACP, pour l'AFC aussi, l'article de Dominique Desbois est très clair et insiste particulièrement sur l'utilisation de la méthode dans SPSS, on peut trouver cet article dans les archives de la revue MODULAD à l'adresse :
http://www-rocq.inria.fr/axis/modulad/archives/numero-18/desbois-18/uneintroduction.pdf

Il est également intéressant de lire sur l'acp et l'afc dans le chapitre consacré aux méthodes factorielles de Tufféry :
http://data.mining.free.fr/cours/Factorielle.pdf

VII) Les séries temporelles
Un document consacré au traitement des séries temporelles par SPSS est développé par Dominique Desbois sous le titre « Une introduction à la méthodologie de Box et Jenkins : l'utilisation de modèles ARIMA avec SPSS », paru dans la revue MODULAD Numéro 18, décembre 2007(p.13-36) et peut être téléchargé à partir de l'adresse suivante :
http://www-rocq.inria.fr/axis/modulad/archivesdetail.htm#18

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc

Youssfi El Kettani - Juillet 2008
youssfielkettani@univ-ibntofail.ac.ma

VIII) Les développements

1) le traitement des cartes géographiques

Le traitement des cartes géographiques par SPSS se présente comme une option supplémentaire. Une présentation de ces possibilités se trouve dans le chapitre 12 de [3] .
On s'aperçoit qu'un certains nombre d'outils complémentaires doivent être intégrés à SPSS, notamment GEOSET MANAGER pour une meilleure exploitation des cartes géographiques. Un autre document plus complet sur les Maps de SPSS est [12].

Par ailleurs, la modélisation statistique des données spatiales est pour le moment absente de SPSS. Cette branche qui trouve ses principales applications en géologie (géostatistique), mais aussi en agronomie et dans les problèmes de l'environnement en général, traite de la modélisation de la variabilité spatiale de variables régionalisées (où les coordonnées dans l'espace sont prises en compte dans l'analyse).
Un document qui présente ce sujet est intitulé « modélisation Géostatistique » par David Causeur, il est à l'adresse : http://www.agrocampus-rennes.fr/math/causeur/PDF/PolyGeostatistique.pdf
Par contre pour le traitement informatique de données spatiales on peut avoir des logiciels libres à l'adresse : http://www.ai-geostats.org/index.php?id=freeware

2) Le data mining et le texte minig

Appelé aussi « fouille des données », le data mining s'est développé pour répondre au foisonnement des méga bases de données qui se sont constituées par accumulation de données dans une multitude d'institutions. Le data mining intègre la globalité des méthodes de traitement et d'exploration, comme en témoigne le site de Tufféry, http://data.mining.free.fr/ . Des documents de synthèse qui présentent bien le sujet intitulés Modélisation statistique et apprentissage ; ainsi que Statistique exploratoire multidimensionnelle sont disponibles sur le site de

Philippe Besse: http://www.math.univ-toulouse.fr/~besse/enseignement.html
Ces techniques de data mining se sont par la suite développées aux variables qui traitent d'objets de façon plus générale au lieu de simples variables réelles uni ou multidimensionnelles. Ainsi il est possible de traiter des chaînes de caractères dans des textes, voire de grands ensembles de textes, ces le texte mining. Le document suivant traite de l'exploration de données textuelles [13]

IX) Conclusion :

Ce survol des ressources pour l'exploration des données par SPSS n'est certainement pas exhaustif. Néanmoins il pourrait faciliter le démarrage et le lancement dans les divers aspects de l'exploration.
Le magazine de la compagnie SPSS inc pour le suivi de l'actualité sur le logiciel est accessible sur le site :
http://www.spss.com/catch/404.cfm?page=http://www.spss.com/nl/spssmagazine/index.htm
On trouve également des sites particulièrement consacrés à la statistique. Le site www.wikistat.ca est dédié aux études de cas en traitement statistique. On y trouve également plusieurs documents pédagogiques de statistique.

Bibliographie :

[1] SPSS Base 14.0 User's Guide 2005 by SPSS Inc. Printed in the U.S.A.
[2] SPSS® 13.0 Brief Guide 2004 by SPSS Inc. Printed in the U.S.A.
[3] SPSS for dummies de Arthur Griffith, 2007, Wiley Publishing, Inc.
[4] SPSS for Beginners, 1999 de Vijay Gupta Published by VJBooks Inc.
[5] A Handbook of Statistical Analyses using SPSS, 2004, de Sabine Landau and Brian S. Everitt, by Chapman & Hall

[6] SPSS for Intermediate Statistics, Use and Interpretation, *Second Edition*
Nancy L. Leech, *University of Colorado at Denver;* Karen C. Barrett , George A. Morgan, *Colorado State University*; In collaboration with Joan Naden Clay, Don Quick
LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS; 2005 Mahwah, New Jersey London
[7] Regression explained , 2000 , Vijay Gupta Published by VJBooks Inc**.**
[8] SPSS Regression Models 12.0, 2003 by SPSS Inc. Printed in the United States of America.
[9] CATEGORICAL DATA ANALYSIS WITH SAS® AND SPSS APPLICATIONS Bayo Lawal, St. Cloud University 2003 LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS Mahwah, New Jersey London
[10] SPSS Categories® 13.0, 2004 by SPSS Inc. Printed in the United States of America.
[11] Tomassone R., Dervin C., Masson J.-P. 1993. *BIOMÉTRIE, Modélisation des phénomènes biologiques*, Masson, Paris, 553 p.
[12] SPSS Maps 10.0, 1999 by SPSS Inc. Printed in the United States of America.
[13] Modélisation probabiliste de langage naturel, 2003, Jardinot et El-Bèse
[14] SPSS for Introductory Stattistics, Morgan & Leech, 2004, LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS Mahwah, New Jersey London

Annexe 1

1	28	130
2	50	280
3	108	650
4	198	800
5	55	268
6	190	790
7	110	500
8	60	320
9	48	250
10	35	250
11	86	350
12	65	300
13	32	155

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc
Youssfi El Kettani - Juillet 2008
youssfielkettani@univ-ibntofail.ac.ma

14	52	245
15	40	200
16	260	1500
17	70	325
18	117	950
19	90	378
20	30	78
21	105	375
22	52	200
23	80	270
24	60	295
25	140	990
26	20	85
27	100	495

.

Annexe 2

Sigle	HCO3	SO4	CL	CA	MG	NA
AIX	341	27	3	84	23	2
BEC	263	23	9	91	5	3
CAY	287	3	5	44	24	23
CHA	298	9	23	96	6	11
CRI	200	15	8	70	2	4
CYR	250	5	20	71	6	11
EVI	357	10	2	78	24	5
FER	311	14	18	73	18	13
HIP	256	6	23	86	3	18
LAU	186	10	16	64	4	9
OGE	183	16	44	48	11	31
OND	398	218	15	157	35	8
PER	348	51	31	140	4	14
RIB	168	24	8	55	5	9
SPA	110	65	5	4	1	3
THO	332	14	8	103	16	5
VER	196	18	6	58	6	13
VIL	59	7	6	16	2	9
VIT	402	306	15	202	36	3

Université Ibn ToFail
Département de Maths et Info
BP 133, Kénitra 14000. Maroc

Youssfi El Kettani - Juillet 2008

youssfielkettani@univ-ibntofail.ac.ma