Tutoriel 3 : Grahiques avec SAS/GRAPH

Résumé

Le module **SAS/GRAPH** *permet de tracer des graphes dont la résolution est adaptée au périphérique utilisé (écran, imprimante, page web. . .) et avec toutes les options possibles. Cette vignette décrit les principales procédures (gchart, gplot, annotate) et options parmi un nombre disponible considérable. Bien que la plupart des procédures statistiques fournissent (*open delevery system*) des graphiques par défaut, la construction spécifique* du *graphique adapté au problème posé est souvent incontournable.*

Plan des tuteuriels :

- *Prise en main*
- *Gestion des données*
- *Graphiques*
- *Macros-commandes*
- *Bases de données*

1 Objectifs

1.1 ODS vs. SAS/Graph

Deux systèmes de production de graphiques sont en "concurrence". Le premier historique est partie intégrante du module SAS/Graph tandis que le deuxième, très récent, est en production depuis la version 9.3 dans le module SAS de base à travers l'ODS (output delivery system) au moment de création de fichiers de sortie au format HTML par défaut et paramétrable en jpeg, png, gif.... La plupart des procédures graphiques apparaissent donc en double : *i.e.* `gplot` de SAS/Graph et `sgplot` de SAS de base, avec évidemment des syntaxes et possibilités différentes mais des principes similaires, comme l'adjonction d'un ensemble d'annotations.

L'intérêt majeur des sorties ODS est de fournir systématiquement et par défaut des sorties graphiques à partir des principales procédures de Statistique (module SAS/Stat) sans nécessité le module SAS/Graph.

Le choix est fait ici de laisser SAS produire les graphiques par défaut dans la plupart des procédures, c'est tout à fait utile dans uns phase exploratoire, mais d'apprendre à utiliser au moins certaines des possibilités offertes par le module SAS/Graph. Celui-ci, très complet, permet en effet de produire la plupart des graphes nécessaires à une illustration pertinente et efficace des résultats statistiques avec une très (trop) large variété d'options. Il serait en effet frustrant de commencer à produire un graphique dans l'environnement ODS, afin de profiter de l'apparente simplicité, pour finalement s'apercevoir que les bonnes options nécessaires à la réalisation du graphique souhaité ne sont disponible que dans SAS/Graph !

1.2 SAS/Graph

Les programmes SAS ci-dessous fournissent les exemples types de graphiques uni et bidimensionnels à partie de SAS/Graph. Certains (histogrammes, profils) ne sont pas une simple utilisation d'une procédure standard, ils nécessitent la construction de tables spécifiques. Il est important de savoir contrôler les options de ses graphiques afin de s'affranchir de celles imposées par la plupart des logiciels guidés par menu. La plupart de ces graphiques sont ceux qui illustrent les vignettes de statistique descriptive élémentaire.

Un soin particulier doit être apporté à la génération de fichiers graphiques. En effet, l'apparente simplicité du copié/collé entre SAS et un traitement de texte sous Windows masque une limitation ; il s'agit d'une simple recopie d'écran qui impose la définition du graphique. En créant un fichier `postscript` ou `jpeg`, on utilise, au moment de l'impression ou de la visualisation, la définition optimale du support. Par ailleurs, la production "programmée " de fichiers graphiques permet d'inclure ceux-ci dans une page `html` ou `xml` pour la mise à jour automatique et en temps réel de sites web.

Des macros seront mises au point lors d'une prochaine séance afin de faciliter l'exécution de ces tâches.

Les petits jeux de données sont saisis au clavier tandis que les plus volumineux sont dans des fichiers accessibles dans le répertoire `data` du siteWikistat.

2 Environnement

2.1 Taille des graphiques

Le graphique demandé est tracé dans une zone dont les dimensions sont définies par les paramètres `hsize` et `vsize` de la commande globale `goptions` (les valeurs maximales sont prises par défaut) diminués de l'espace nécessaire à l'édition des titres, sous-titres, notes, légendes,.... Les dimensions peuvent être exprimées en trois unités : `pouce, cm` ou `pct` qui signifie "pourcentage de la dimension totale". Cette dernière unité est préférable pour exprimer les tailles de caractères et symboles lorsque les dimensions globales, liés au périphérique de sortie, sont sujettes à modifications.

2.2 Sauvegarde des graphiques

En l'absence de commande explicite, le graphe apparaît sur le périphérique par défaut, l'écran de l'ordinateur, dans la fenêtre de visualisation des résultats. Une fois les graphiques mis au point, ils peuvent être sauvés dans des fichiers ou un traitement de texte.

Un fichier de format `.jpg` est systématiquement créé (sous windows) dans un répertoire temporaire ainsi qu'un fichier `sashtml.html` contenant le graphique ; consulter la fenêtre du journal. Il devrait être possible de contrôler la destination de ce fichier.

2.3 Image écran

Un clic droit sur le graphique (windows mais unix ?) ouvre un menu qui permet de sauver l'image dans le format `.png` avec la définition de l'écran.

2.4 ODS

Comme vu en introduction, le graphique peut être automatiquement orienté dans une fichier au bon format (.rtf, .html...) et sans doute avec une meilleure définition que celle de l'écran.

```
ODS RTF BODY='nomfichier.rtf';
ODS GRAPHICS ON;
/* Programme SAS */
ODS GRAPHICS OFF;
ODS RTF CLOSE;
```

3 Commandes globales

Elles définissent des objets (axes, symboles, trames, légendes) et les options utilisés pour les tracés ; elles demeurent valables jusqu'à une nouvelle définition ou la fin de la session sas.

3.1 Axes

Des types d'axes, numérotés de 1 à 99 sont définis avant de pouvoir être utilisés dans les différents graphiques. Ils précisent l'échelle (liste de valeurs, logarithmique), l'apparence (longueur, couleur, épaisseur, style de ligne, origine), les marques d'échelle (nombre, couleur, épaisseur, hauteur), les valeurs des échelles (format), le libellé (police,...).

```
axis1 order=(1973 to 1981 by 2)
   label=('annee')
   minor=(number=1)
   width=3;
axis2 order=(0 to 10000 by 1000)
   label=('Revenu en francs')
   minor=none
   width=3;
```

3.2 Légendes

Comme pour les axes, différents types de légendes (de 1 à 99) sont définissables. Ils spécifient positions et textes des libellés qui identifient les différents graphismes et symboles utilisés.

3.3 Symboles

Les différents types de symboles (1 à 99) sont définis afin de décrire les modes de représentation recherchés. Sont concernés : le symbole (forme, taille, couleur) utilisé pour représenter un point, le type de lignes reliant les points (couleur, continue, hachurée, pointillée,...), la façon ou mode d'interpolation incluant barres, boîtes à moustaches, escaliers, splines, intervalles de

confiance, régression (linéaire, polynomiale, spline).

```
symbol1 interpol=sm50s  /* lissage spline */
    value=diamond      /* symbole        */
    height=3           /* taille du symb.*/
    width=2;           /* epaisseur      */
```

3.4 Options graphiques

Outre ceux décrits ci-dessus (`hsize, vsize`), cette commande redéfinit les valeurs de plus de 80 paramètres affectant

- les différents aspects du graphique :
  - *border* cadre autour du graphique,
  - *gunit=cm|in|pct* unité de mesure,
  - *rotate=landscape|portrait* orientation du graphique,
- le texte :
  - *ftext* police du texte,
  - *ftitle* police des titres,
- texte, symboles, types de hachures, légendes.
- Les paramètres reprennent leurs valeurs par défaut à la suite de :
  - *reset=all|global all* concerne tous les paramètres tandis que que *global* n'affecte pas ceux définis dans la même commande.

3.5 Titres et notes

Les commandes `title` et `footnote` définissent des lignes de texte autour du graphique, elles suivent le même principe que celui décrit au paragraphe I.1.4 et d'autres options sont disponibles : taille, couleur et police des caractères, position, rotations de la ligne de texte et des caractéres, tracés de lignes.

```
goptions reset=global gunit=pct border
    ftext=swissb htext=3;
title1 height=5 'Institut';
title2          'de';
title3 height=5 'Mathématiques';
footnote1 font=script justify=left
    'Universite de Toulouse';
```

Il est important de noter que chaque paramètre peut être initialisé ou redéfini à différents endroits d'un programme SAS : dans les commandes spécifiques (`symbol, legend, axes, pattern, title, footnote`), par la commande `goptions` et dans chacune des procédures. Ceci impose de bien distinguer les paramètres globaux, applicables à tous les graphes, des paramètres spécifiques à chaque graphe.

Réaliser les graphes suivants en exécutant les commande. Se reporter à l'annexe pour expliciter et comprendre la syntaxe de chaque procédure.

4 Quantitatif

Graphes usuels pour des variables quantitatives discrètes puis continues.

4.1 Discret

Création de la table

Lire les données de répartition en âge, à chaque âge est associé un effectif.

```
data age;
input age effectif eff_cum;
cards;
24 1 1
26 2 3
29 3 6
31 2 8
33 4 12
37 2 14
38 4 18
41 3 21
43 3 24
45 1 25
46 6 31
49 3 34
50 1 35
52 3 38
57 5 43
59 2 45
60 2 47
```

```
62 1 48
;
/* place impérative du ";" */
run;
proc print;run;
```

Diagramme en bâton

Le graphe légitime est un diagramme bâton, pas un histogramme qui correspond à une variable continue.

```
proc gplot data=age;
axis1 label=("Age" justify=right);
axis2 label=("Effectif")
      order=(0 to 6 by 1) offset=(0,);
symbol1 interpol=needle value=dot;
plot effectif*age / haxis=axis1
     vaxis=axis2 hminor=4 vminor=0;
run;
quit;
```

De même, comme la variables est discrète, le diagramme cumulatif qui estime la fonciton de répartition n'est pas continu.

Diagramme cumulatif

```
proc gplot data=age;
axis1 label=("Age" justify=right)
      order=(20 to 70 by 10);
axis2 label=("Effectif" justify=right "Cumule")
      order=(0 to 50 by 10) offset=(0,);
symbol1 interpol=step  value=none;
plot eff_cum*age /
      haxis=axis1 vaxis=axis2
      hminor=4 vminor=0;
run;
quit;
```

4.2 Continu

Répartition des exploitations agricoles par classes de surface agricole utile (SAU). Repérer comment sont organisées les données, attention, les classes ne sont pas d'amplitudes égales.

Fonction de répartition

```
data exploit;
input SAU dens_eff freq_cum;
cards;
0   0     0
5   4.8   0.24
10  2.18  0.349
20  1.78  0.527
35  1.35  0.73
50  0.68  0.832
200 0.112 1.
;
run;
proc print;run;

proc gplot data=exploit;
axis1 label=("SAU(ha)" justify=right)
      order=(0 to  200 by 50) ;
axis2 label=("frequences" justify=right "Cumulees")
      order=(0 to 1 by 0.5) offset=(0,);
symbol1 interpol=join value=dot;
plot freq_cum*sau / haxis=axis1 vaxis=axis2
     hminor=0 vminor=4;
run;
quit;
```

Histogramme

```
proc gplot data=exploit;
axis1 label=("SAU(ha)" justify=right)
      order=(0 to  200 by 50);
axis2 label=("densite de" justify=right "frequence")
      length=6cm order=(0 to 5 by 1) offset=(0,);
 pattern1 value=msolid ;
```

```
symbol1 interpol=steprj value=none;
plot dens_eff*sau / haxis=axis1 vaxis=axis2
      hminor=4 vminor=0 areas=1;
run;
goptions reset=all;
quit;
```

5 Qualitatif

Graphiques destinées aux variables qualitatives.

5.1 Création de la table

Répartition des catégories socio-professionnelles parmi les actifs.

```
data csp;
input csp $ 1-10  eff classe;
cards;
ar et com  1739  1
agricult   1312  1
cadres     2267  1
p  inter   4327  1
employes   5815  1
ouvriers   6049  1
;
run; proc print;
```

5.2 Diagramme en barres

```
axis2  label=("Effectif")
       order=(0 to 8000 by 2000);
proc gchart data=csp;
 vbar csp / sumvar=eff  raxis=axis2
          vminor=1;
run;
quit;
goptions reset=all;
```

Améliorer l'axe des ordonnées en changeant la commande `order`.

5.3 Diagramme en colonne

Les graphiques sont en couleur par défaut :

```
axis3 label=none ;
axis4 length=1cm;
proc gchart data=csp;
hbar classe / noaxis subgroup=csp raxis=axis3
           gaxis=axis4   sumvar=eff nostats;
run;
quit;
goptions reset=all;
```

Graphique en noir pour impression avec hachurage automatique :

```
goptions colors=(black);
axis3 label=none ;
axis4 length=1cm;
proc gchart data=csp;
hbar classe / noaxis subgroup=csp raxis=axis3
           gaxis=axis4   sumvar=eff nostats;
run;
quit;
goptions reset=all;
```

5.4 Diagramme en secteur

Le graphique est vide par défaut. on peut lui rajouter des hachures :

```
pattern1 v=p2n0 ;
pattern5 v=p2n90;
pattern3 v=P2x45;
pattern4 v=P4n0;
pattern2 v=P4n90;
pattern6 v=P4x90;
goptions colors=(black);
proc gchart data=csp;
pie csp /sumvar=eff noheading  slice=outside
            percent=outside;
run;
quit;
goptions reset=all;
```

ou des couleurs :

```
pattern1 v=solid;
proc gchart data=csp;
pie csp /sumvar=eff noheading  slice=outside
             percent=outside;
run;
quit;
goptions reset=all;
```

Un peu plus de travail permettrait de sélectionner des couleurs plus équilibrées. Il est évidemment possible mais pas recommandé de remplacer `pie` par `pie3d`.

6 Bidimensionnel

6.1 Boîtes parallèles

Les données sont issues d'une expérience clinique. Le rythme cardiaque est mesuré en fonction de la concentration (facteur à 3 niveaux) d'une molécule en cours de test.

```
data pento;
infile "pento.dat";
input code $ rythme facteur $;
run;
proc print;run;

proc gplot data=pento;
symbol1 interpol=box;
axis1 length=5cm offset=(1cm,1cm) ;
axis2 length=5cm;
plot  rythme*facteur=1/vaxis=axis2 vminor=1
            haxis=axis1 ;
run;
goptions reset=all;
quit;
```

La procédure `boxplot` permet de contrôler les paramètres graphiques des boîtes avec de très nombreuses options.

```
proc boxplot data=pento;

axis1 length=5cm  ;
axis2 length=7cm;
plot  rythme*facteur/notches vaxis=axis2 vminor=1
           haxis=axis1 ;
run;
goptions reset=all;
quit;
```

6.2 Profils

Voir aussi les graphiques de type "mosaïque" dans sas/insight. Les données décrivent l'âge d'obtention du bac, la durée pour obtenir le DEUG (L2) et un pourcentage calculé pour représenter des profils. Il est dans ce cas évidemment plus simple d'utilise rle mosaÏc plot de SAS/Insght.

```
data deug;
infile  "profdeug.dat";
input age $ duree $ prof;
run;
proc print;run;
axis1 length=5cm  label=("Age au BAC");
axis2 length=5cm  label=("Pourcentage") /*
          order=(0 to 100 by 20)*/;
goptions colors=(black);
proc gchart data=deug;
vbar age / subgroup=duree sumvar=prof
        maxis=axis1  raxis=axis2 vminor=1
        midpoints="moyen" "inf18" "18ans"
                          "19ans" "sup19";
run;
quit;
goptions reset=all;
```

Refaire le même graphique en couleur.

6.3 Nuage

Il s'agit de représenter l'évolution du chiffre d'affaire des entreprises en fonction du nombre de salarié.

```
data entr;
infile "entr.dat";
input code $ nb ef ca;
run;
proc print;run;

proc gplot data=entr;
symbol1 interpol=r value=dot;
axis1 label=("Chiffre d affaire" justify=right);
axis2 label=("Effectif")  offset=(0,);
plot ca*ef=1 / haxis=axis1 vaxis=axis2;
run;
quit;
goptions reset=all;
```

7 Énergies renouvelables

Le fichier de données `energie.txt` contient la production d'énergie d'origine renouvelable en France de 2001 à 2003, exprimée en Gwh. Les différentes sources sont l'énergie hydraulique, solaire, éolienne, l'énergie issue des déchets urbains solides, du bois et des déchets du bois, les biogaz (source ; Ministère de l'Économie, des Finances et de l'Industrie, 2004). On dispose donc en tout de trois variables `annee`, `typeen` et `prod`. Exécuter et commenter le code ci-dessous le plus précisément possible :

```
data enprod;
infile "energie.txt";
input annee typeen $ prod;
proc print;run;

goptions reset=global vpos=45
htitle=2 htext=1 hpos=100;
title1 "Production d energie 2001 a 2003";
proc gchart data=enprod;
hbar3d typeen / sumvar=prod type=sum group=annee;
run;
quit;
```

A quoi sert l'option `group=` ? Remplacez `type=sum` par `type=mean`. Essayer l'option `patternid=group` placée après `hbar typeen`. Tester de même les options `gspace=5`, `noaxos`, `nostats`, `descending`.

Création d'un graphe représentant la production cumulée d'énergie pour chaque année :

```
goptions reset=global cback=white htitle=3
    htext=1 hpos=100 vpos=45;
title1 "Production d""énergie 2001 à 2003";
pattern1 value=solid color=yellow;
pattern2 value=solid color=blue;
pattern3 value=x3 color=green;
pattern4 value=solid color=red;
pattern5 value=x2 color=blue;
pattern6 value=solid color=green;
axis1 label=("Energie produite (*)");
footnote1 justify=left "(*) exprimee en Kwh"
justify=right;
axis2 label=none;
legend1 label=(position=(topo left)
              "Type d" justify=left "energie")
value=("biogaz" "bois" "dechets urbains solides"
              "eolienne" "hydraulique" "solaire");
proc gchart data=enprod;
vbar annee / sumvar=prod sum discrete raxis=axis1
maxis=axis2 space=3 width=7
subgroup=typeen legend=legend1;
run;
quit;
goptions reset=all;
```

A quoi sert l'option `sum` ?

L'énergie hydraulique prend trop de place, celle-ci est retirée du graphique.

```
data enprod2;
set enprod;
if typeen="hydrau" then delete;
```

```
run;
proc format;
value $typeen "biogaz"="Biogaz" "eolien"="Eolienne"
"sol"="Solaire" "dechsol"="Dechets solides"
   "bois"="Bois";
run;
proc gchart date=enprod2;
pie typeen / sumvar=prod type=sum other=5
coutline=black noheading group=annee across=2;
format typeen $typeen.;
where annee in(2001 2003);
run;
goptions reset=all;
```

Annexes : Syntaxe des commandes

procédure “gchart”

Cette procédure trace des diagrammes en barres (`hbar`), en colonnes et histogrammes (`vbar`), en secteurs (`pie`) et aréolaires (`star`). Elle peut traiter des variables quantitatives ou qualitatives ; les variables quantitatives sont codées explicitement ou automatiquement en classes ou, selon les besoins, sommées ou moyennées.

Syntaxe

proc gchart <options générales> ;
by <descending> variable ;
vbar liste de variables
</<options d’apparence>
<options statistiques> <options d’axes> > > ;
hbar liste de variables
</<options d’apparence>
<options statistiques>
<options d’axes> > ;
pie liste de variables
</<options d’apparence>
<options statistiques> ;
star liste de variables
</<options d’apparence>
<options statistiques> > ;

Options générales

- *data=table sas* indique le nom de la table ou, par défaut, la dernière crée,
- *annotate=* table contenant les compléments graphiques.

Options d’apparence

Elles spécifient les couleurs, les espacements et largeurs de colonnes ou barres. Il est également possible d’adjoindre un cadre (`frame`), de supprimer (`nolegend`) ou modifier la légende. Une option `annotate` peut être introduite au niveau de chaque commande.

Options statistiques

- *sumvar=* variable quantitative dont le cumul ou la moyenne est représenté,
- *freq=* variable de pondération des observations,
- *midpoints=* liste des bornes de classes,
- *levels=* nombre de classes,
- *type=* spécifie ce que représente le graphique (par défaut une fréquence) : `cfreq` (fréquence cumulée), `cpt` (pourcentage cumulé), `pct` (pourcentage), `sum` ou `mean` (associées à `sumvar=`).
- *group=* représentation de plusieurs graphes côte à côte suivant les modalités de la variable spécifiée (`hbar` ou `vbar`),
- *subgroup=* découpage des barres ou colonnes selon la participation des modalités de la variable spécifiée (`hbar` ou `vbar`).

Options d’axes

Deux options permettent de de définir les axes ou de leur assigner des déclarations antérieurs : `gaxis=axis`n pour l’axe des groupes et `maxis=axis`n pour celui des bornes où n caractérise la définition d’axe concernée (cf. paragraphe V.2.1.).

procédure “gplot”

Graphiques en haute résolution de nuages de points en deux dimensions.

Syntaxe

proc gplot <options générales> ;
by <descending> variable ;
plot liste de graphiques
</ < *annotate*=data-set >
< options d'apparence>
< options d'axes>> ;
bubble liste de graphiques
</ < *annotate*=data-set >
< options d'apparence>
< options d'axes>> ;

Options générales

- *data=table sas* indique le nom de la table ou, par défaut, la dernière crée,
- *annotate=table sas* table contenant les compléments graphiques.
- *uniform* impose les mêmes échelles aux axes des différents graphiques.

Options d'apparence

Elles spécifient les couleurs, les polices de caractères, les tailles des bulles (*bubble*), le hachurage d'aires, la définition de légendes, la superposition (`plot`).

Options d'axes

Deux options permettent de définir les axes ou de leur assigner des déclarations antérieurs : `vaxis=axis`n pour l'axe vertical et `haxis=axis`n pour l'axe horizontal où n caractérise la définition d'axe concernée (cf. paragraphe 5.2.1.). De plus, `frame` trace un cadre tandis que `noaxis` supprime les axes.

Commandes

by suivi du nom d'une variable qualitative indique que les graphiques sont tracés par groupe d'observations ; la table doit être triée.

plot liste des graphes sous la forme : `y*x<=`n`|variable>`, avec la même syntaxe que précédemment pour désigner plusieurs graphes (a*(a b),...). La variable y fournit les ordonnées et x les abscisses des points représentés par des symboles définis dans la commande `symbol`n ou par différents symboles selon les valeurs de la `variable` spécifiée qui induit une classification. Dans ce dernier cas, une légende est créée par défaut.

bubble liste des graphes sous la forme : `y*x = size` où `size` est une variable indiquant la taille des bulles à tracer autour des centres de coordonnées x et y.

Annotate data set

Une *table d'annotations*, définie lors d'une étape data, est une table SAS contenant les descriptifs d'un ensembles de graphiques qui viendront se superposer aux résultats des procédures précédemment décrites (`gchart, gplot, ...`). Il est alors possible de positionner tout libellé ou toute figure géométrique simple et ainsi de personnaliser ses graphiques.

Des applications immédiates sont, par exemple, la production de plans factoriels avec identifications des points (variables, individus, modalités) par des libellés explicites ou encore le tracé du cercle des corrélations en analyse en composantes principales.

Par principe, chaque ligne ou “observation” d'une table d'annotations est une commande de réalisation d'un graphique particulier. Les valeurs de chacune des “variables” spécifient comment réaliser ce graphique : type, emplacement, couleur,.... Les variables de la table d'annotation ont des noms prédéfinis ; les plus usuelles sont :

function indique ce qu'il faut tracé : *bar, draw, frame, pie, symbol, label,...*,

x positionnement en abscisses,

y positionnement en ordonnées,

size hauteur des caractères,

xsys unité de mesure des abscisses,

ysys unité de mesure des ordonnées,

hsys unité de mesure des hauteurs,

color couleur,

position d'un texte par rapport aux coordonnées (calé à gauche, centré,...),

line type de ligne (par défaut, continue),

text texte du libellé

style police de caractères.

La mise en œuvre de ces fonctionnalités est un peu fastidieuse mais c'est la seule façon de faire éditer par SAS certains types de graphiques dont les fameux plans factoriels avec les libellés en clair de tous les points.

Création d'une table d'annotations :

```
data annocomp;
  set outcomp;
  x    = prin1;
  y    = prin2;
  xsys= '2';
  ysys= '3';
  text= lib_ind;
  size= 0.8;
  label x = 'axe1';
  label y = 'axe2';
  keep x y text xsys ysys size;
run;
```