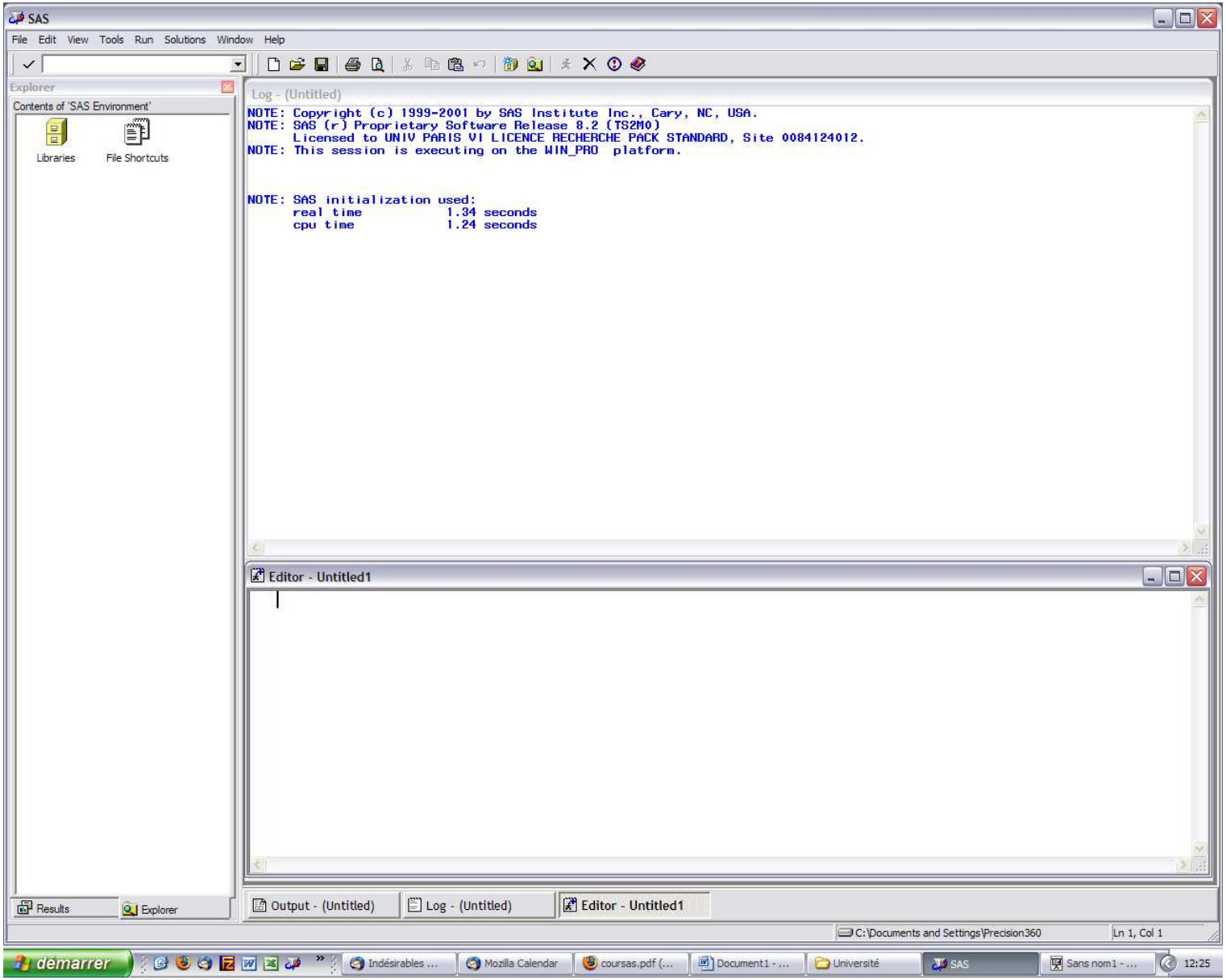
Introduction au logiciel SAS

Olivier Godechot

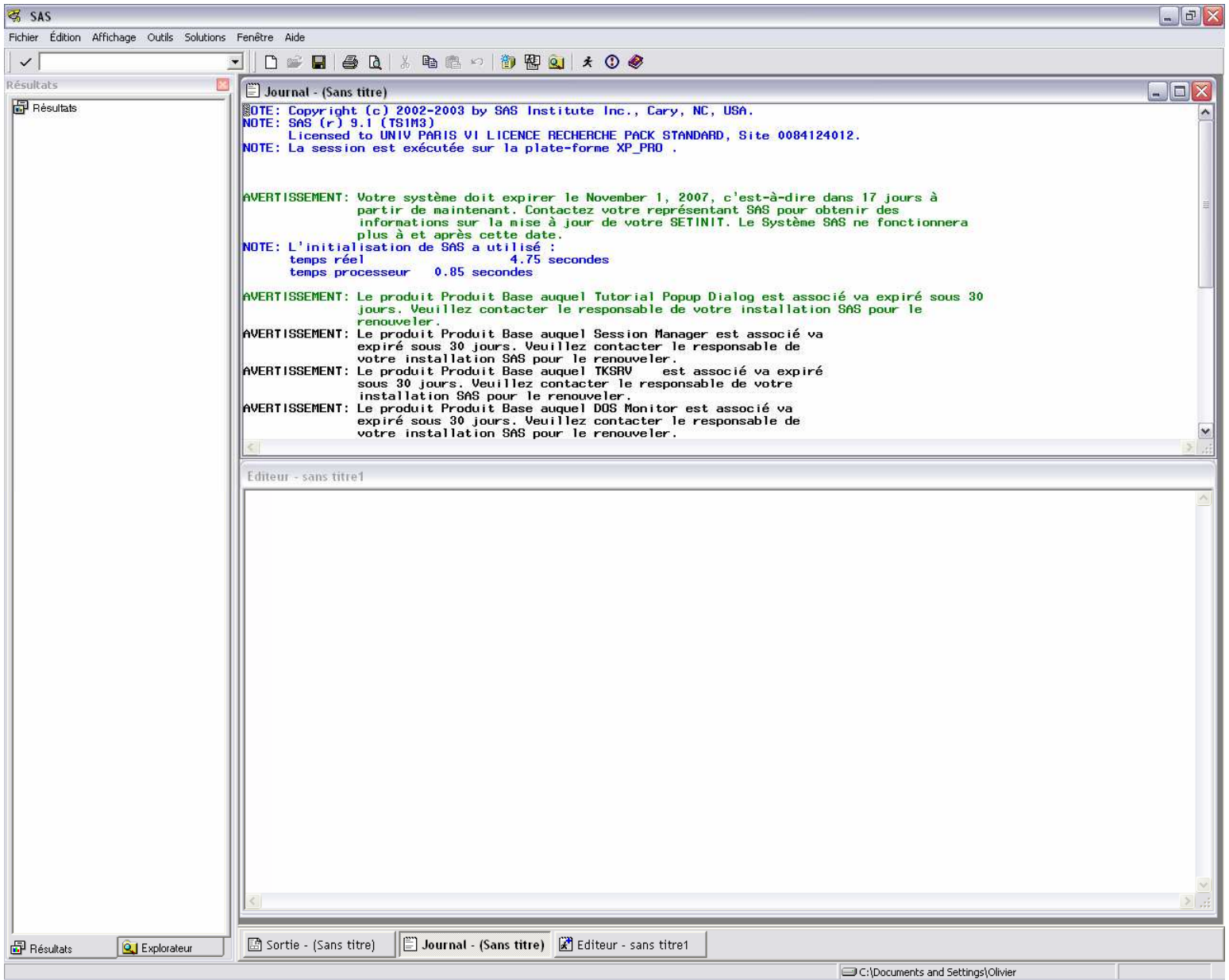
SAS (9.13). Plan d'attaque

- À quoi ressemble le logiciel ?
- Manipuler les données (étape data)
- Quelques procédures statistiques de base (étape proc)
- Fusion des données
- Import et export des données

À quoi ça ressemble



Le même en version française (9.13)



Les fenêtres

- Fenêtre *Program Editor [Editeur]* ou *Enhanced Program Editor [Editeur amélioré]* : écriture du programme
  - *Enhanced Program Editor sous SAS 8 et supérieur, écriture colorée, très pratique pour repérer les erreurs.*
- Fenêtre *Log [Journal] :* conséquence de la soumission du programme.
  - *En bleu bon déroulement. En vert, anomalie. En rouge, erreur.*
- Fenêtre *Output [Sortie] :* résultats des procédures statistiques
- Fenêtre *Results [Résultats].*
  - Sommaire de la fenêtre *Output [Sortie]*
  - Accès aux résultats au format html
- Fenêtre *Explorer [Explorateur]*
  - Visualisation et manipulation des bases de données.

Aperçu des menus déroulants

- *File [Fichier] :* gestion des fichiers (mais les fichiers de programme)
- *Edit [Edition] :* Annulation, Copier/Coller, Trouver, Sélection
- *View [Affichage] :* Génération des fenêtres
- *Tools [Outils] :* Je ne m'en sers pas... (sauf préférences)
- *Run [Executer] :* Soumettre un programme
- *Solutions :* Module de programmes clefs en main
- *Window [Fenêtre] :* Sélection des fenêtres
- *Help [Aide] :* Aide (important).
  - *http://v8doc.sas.com/sashtml/*
  - *http://support.sas.com/onlinedoc/913/docMainpage.jsp*

La grammaire d'un programme

- Un même mot -> plusieurs statuts grammaticaux
- **Instructions :** Une instruction se termine toujours par un point virgule

```
data a; * Créer la table a;
```

- **Options :** Un sous élément optionnel d'une instruction

```
data a (keep= x ); * Créer la table a et ne
  garder que la variable x;
```

- **Fonction :** une opération sur une variable

```
y=max(k,l,m);*y est le maximum des variables k, l, m;
```

- **Procédure :** Un programme statistique tout fait composé de plusieurs instructions, commençant par *proc et se terminant par run; ou par quit;*

```
proc print data=a; run;
```

Étape pour faire un programme

- Étape 1: Le Libname. Indication à SAS de la localisation des bases.
- Étape 2 : écriture du programme SAS
  - Étapes DATA : construction, sélection des bases de données, créations de variables
  - Étapes Proc : opérations statistiques sur les variables. Tableaux croisés, moyennes...
- Étape 3 : Soumission du programme, ou d'une partie de programme
- Étape 4 : Vérification de la *log (pour voir s'il y a des erreurs)*
- Étape 5 : Voir les résultats lst.
- SAS Manipule les données à distance !

Le `libname`

- SAS mal intégré à Windows. Mauvaise connaissance des répertoires.
- On est obligé de lui indiquer les répertoires dans la syntaxe SAS. C'est le libname

```
libname toto "d:\O_Godechot\StrucSal\strucsal1992\";
```

- Le fichier **monfichier.sd2 ou monfichier.sas7bdat** contenu dans le répertoire s'appelle alors **toto.monfichier** pour SAS
- On peut visualiser alors ce fichier

TP 1. Libname + visualisation d'un fichier

Étape `data`

- Fichier INSEE. Se munir du questionnaire et du dictionnaire des codes.
- Le cas échéant vérifier le contenu de la base

```
proc contents data=toto.t3; run;
```

- Premier principe, ne jamais modifier directement une table originale. Créer une table temporaire. (économie de place + sécurité)

```
data a; set toto.t3;
run;
```

- Je crée une table temporaire appelée a. J'y verse le fichier toto.t3

Création / Modification de variables

- Les opérations se font ligne par ligne. On additionne les colonnes d'une même ligne (et non les lignes d'une même colonne).
- Intérêt de cette étape data ➔ création de variables
- Toutes les opérations mathématiques
  - Opérations élémentaires : + * - /
    - Var3=var1+var2; Var4=var1/var3;
  - Puissance : **
    - Var5=var1**2;
  - Logarithme : var6=log(var1);
  - Probabilité : Var7=Probnorm(var1);
  - Statistiques : Var8=std(var1,var2,var3,var4);
- Opérations sur du texte

```
z=count("abba", "ab");
```

- Opérateurs booléens :
  - **if then else do NE and or**

```
data a; set toto.t3;
if s="1" then sexe="homme";
else if s="2" then sexe="femme";
run;
```

Variables numériques / variables caractères

- Pour gagner de l'espace disque, l'INSEE code en général les variables numériques en mode caractère
- Pour pouvoir faire des procédures numériques (proc means, proc reg), nécessité de transformer les variables caractères en variables numériques

```
nempn=NEMP*1;
```

- Il est important de savoir si on travaille sur du numérique ou du caractère. En numérique pas de guillemets. En caractère, mettre des guillemets.
- Dans certains cas, on peut vouloir transformer du numérique en caractère

```
nempnc=nempn !! "" ;
```

Les Non-réponses

- Non-réponses
  - Une donnée manquante en variable caractère

```
if NEMP="" then NEMPBIS="Non réponse";
```

  - Une donnée manquante en variable numérique:

```
if NEMPN=. then NEMPTERS="Non réponse";
```

  - Les non-réponses à l'INSEE.
    - Parfois, distinction des non-réponses et des non-questions (questions non-posées). Codage des non-réponses avec des 9. Autant de 9 que de nombre caractères pour la variable.

```
if salair="999999" then salairn=.;
```

Créations de variables (fin)

- Les variables dichotomiques (très pratiques pour les régressions) Vaut 1 si le cas se réalise.

```
sexe1=(s="1");
sexe2=(s="2");
label sexe1="homme";
label sexe2="femme";
```

- Les labels. Un descriptif des variables
- Une combinaison de plusieurs méthodes :

```
pfix=((10<NCOMP1<15)*RCOMP1+(10<NCOMP2<15)*RCOMP2+(10<NCOMP3<15)*RCOMP3
+(10<NCOMP4<15)*RCOMP4+(10<NCOMP5<15)*RCOMP5);
```

Il y a toujours plusieurs écritures possibles pour un même résultat !

Étape Data. Évolution SAS 6-SAS 9.13

- La taille maximum des variables caractères est passée de 200 caractères à 32767 caractères.
- La taille des noms de variables passée de 8 à 32 caractères, des labels de 40 à 256 caractères.
- Distinction majuscule, minuscule dans les noms de variable.
- Nouvelles fonctions, nouvelles procédures

L'étape Proc / Quelques proc utiles

- Statistiques descriptives
  - `Proc freq` (tableau croisé, fréquence)
  - `Proc means` (moyennes)
  - `Proc univariate` (fractiles)
  - `Proc tabulate` (tableaux croisés à entrées multiples)
- Analyse factorielle
  - `Proc corresp`
- Régression
  - `Proc reg` (MCO)
  - `Proc logistic` (régression logistique)
- Utilitaire
  - `Proc contents` (contenu)
  - `Proc print` (imprimer la table)
  - `Proc plot` (graphique)
  - `Proc Sort` (Tri)
  - `Proc format` (étiquette pour tableau)

Proc freq

```
data a; set toto.t3;
if s="1" then sexe="homme";
else if s="2" then sexe="femme";
label dmgt1="Déménagement pour nouvel emploi";
run;

proc format;
value $ouinon
"1"="Oui"
"2"="Non";
run;

proc freq data=a;
tables sexe*dmgt1/chisq;
format dmgt1 $ouinon.;
run;
```

Résultat

Table of sexe by DMGT1

sexe / Frequency / Percent / Row Pct / Col Pct	DMGT1(Déménagement pour nouvel emploi) Oui	Non	Total
femme	597	2930	3527
	7.21	35.37	42.57
	16.93	83.07	
	30.92	46.11	
homme	1334	3424	4758
	16.10	41.33	57.43
	28.04	71.96	
	69.08	53.89	
Total	1931	6354	8285
	23.31	76.69	100.00

Frequency Missing = 1855

TP

- Genre et diverses formes de mobilité (NMETIER, DMGT2 – DMGT5)
- Différence de mobilité par genre chez les riches et les pauvres.
- Moyenne de changements d'emploi, par genre, par genre et par revenu.
    - Quelques aides : instruction `where` permet de spécifier un domaine de validité dans les étapes proc

      ```
      where an>"55";
      ```

    - Proc means :

      ```
      proc means data=a; var nempn; class s; run;
      ```

Fusion de fichiers – Merge

- Un des apports de SAS par rapport à Excel → gestion de la fusion
  - Instruction merge
  - Domaine de validité : une des deux tables au moins ne doit avoir qu'une seule ligne par identifiant.
  - Données plus complexes => Proc sql;
- Ici quatre tables
- On peut vouloir fusionner les deux tables salariés

  - Clé de fusion identique dans les deux tables

  - Trier selon les clés de fusions.

```
proc sort data=a; by ident; run;
```

  - Fusion

```
data c; merge a b; by ident; run;
```

Exportation sous Excel/Word

- Exportation des résultats
  - Méthode 1 à partir de l'output (Sortie) :
    - 1.a. Copier / coller sous word
    - 1.b Enregistrement de la lst sous format .rtf ou sous format .lst
    - 1.c Copier / coller sous Excel + Ensuite Menus Données/Convertir/ Largeur fixe et placer les barres.
  - Méthode 2 à partir des sorties HTML
    - Le cas échéant : Outils/Options/Préférences/Résultats Cocher Créer Html (si ce n'est pas déjà fait) Mise en page (minimale)
    - Dérouler le menu résultats. Voir le fichier html
      - Copier Coller sous word
      - Copier coller sous excel
- Exportation des données sous excel
  - Menu Fichier Exporter les données

    Format excel échoue souvent. Format DBF ou Txt a un meilleur taux de succès
  - Menu Explorateur.
    - Sélectionner les tables View in Excel

Créer / Importer des données

- Plusieurs méthodes
  - Cards (à n'utiliser que pour des exemples très simples)

```
data z;
input sexe $ age;
cards;
homme 15
femme 12
homme 08
ns .
homme .
.  10
;
run;
```

  - Menu Fichier / import

Organisation des données : noms de variables sur la première ligne. Attention aux erreurs possibles. Format de dates. Mélange caractères/chiffres, etc. Bonne gestion des conversions de décimales.

Importation des données avec le menu automatique

- La difficulté c'est la reconnaissance des variables comme variables caractères ou numériques
- SAS scanne les 8 premières lignes des variables excel
  - Si c'est majoritairement du texte → caractère
  - Si c'est majoritairement des chiffres → numérique
- Format difficultés à l'importation
  - Reconnaissance de la taille maximale des variables si la taille max du contenu < 255.

Autres méthodes d'importation

```
filename essai dde
"excel|d:\O_Godechot\[DonneesAtelier.xls]BASERECODEE!l2c1:l92c204"
 notab lrecl=300000;

data mabase;
infile essai dlm='09'x notab dsd;
length
     IDENT $5
     SEXE $1
     ANAIS $4
     /*etc*/
     COMMENT $200
;
input
     IDENT $
     SEXE $
     ANAIS $
     /*etc*/
     COMMENT $
;
run;
```

Avantages et inconvénients de SAS

- Inconvénients de SAS
  - Coût d'apprentissage
  - Coût du logiciel/disponibilité
- Avantages
  - Rapidité (on n'est pas obligé de refaire 10 clics pour refaire un programme)
  - Sûreté (distinction base de données/modifications)
  - Souplesse (grande possibilité de manipulations de données)
  - Puissance (Un grand nombre de procédures statistiques)

Pour ne pas finir

- Cette introduction n'a de sens que branchée sur une pratique !!!
- Sans application, elle sera stérile