Les opportunités du data mining pour les statistiques publiques

Une application des techniques d'intelligence artificielle aux données de l'Enquête suisse sur la population active (ESPA)

Eric Stephani, Sandro Petrillo, Laura Azzimonti°, Mauro Scanagatta°, Marco Zaffalon°

USTAT, °IDSIA/Supsi

Ittingen, 20 novembre 2017

Introduction du projet

Quand? | Bref histoire du projet

1. Participation à un cours de Data mining

2. Volonté partagée d'essayer une application des techniques d'intelligence artificielle avec une BD tirée de la statistique publique
   - Choix et acheminement du projet
   - Premiers résultats
   - Avancement selon la méthode du tâtonnement

3. Partecipation aux Journées Suisse de la statistique, **opportunité de discussion et d' échange**.

Qui? | Acteurs du projet

- **USTAT**, Office de statistique du Canton du Tessin
- **IDSIA**, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (SUPSI, Département technologies innovantes)

- **OFS**, Office fédéral de la statistique

Quoi? | Choix de travailler avec l'ESPA

- L'IDSIA nous a demandé une source de données avec beaucoup de variables et pas, nécessairement, beaucoup d'observations

- Les résultats de la statistique ESPA (Enquête suisse sur la population active) font confiance à une approche "traditionnelle" pour l'estimation de la population de référence et des sous-groupes (ex.: chômeurs, hommes, suisses ou étrangers,...).

- L'estimation est indiquée comme incertaine, seulement si:
  - l'échantillon est relativement petit (n $<$ 90 observations),
  - ou même non publiée quand l'échantillon est réduit à très peu de cas (n $<$ 5 observations).
    Dans ces cas, les éstimations sont publiées entre parenthèses dans les tableaux de l'OFS ou sont remplacées par un "X".

- Existe-t-il d'**autres approches** ou méthodes pour estimer des statistiques?

Comment? | Les réseaux bayésiens (discrets)

- En gros, un réseau bayésien (discret) est:
  - Une façon de représenter la distribution conjointe d'un ensemble de variables
- Un réseau bayésien est défini par le biais de:
  - Un **graphe orienté acyclique** (DAG: directed acyclical graph) entre les variables qui indique entre quelles variables (et dans quel sens) se situent les dépendances directes
  - Une distribution de probabilité *locale* pour chacune des variables:
    - une distribution conditionnelle pour les variables qui dépendent d'autres variables
    - une distribution marginale pour les variables restantes

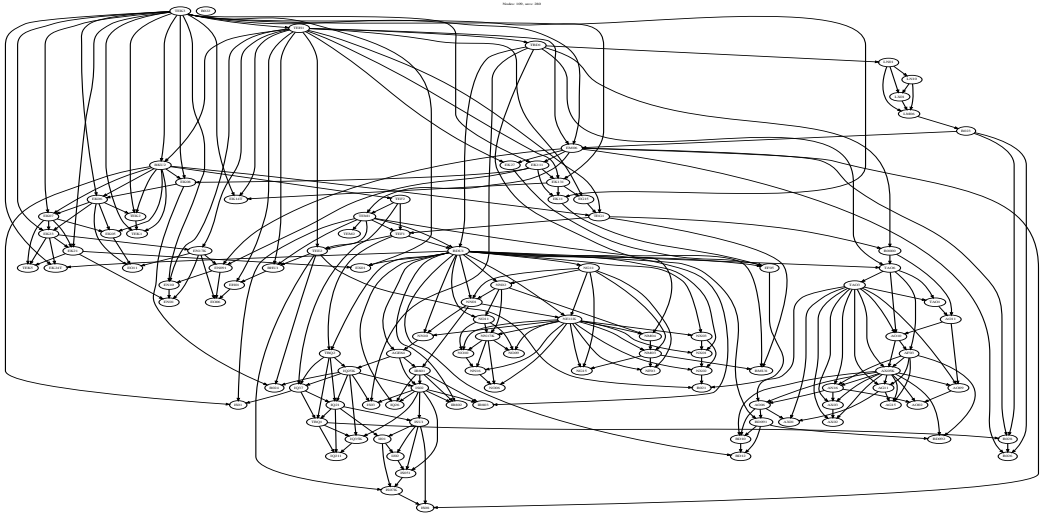
Un réseau bayésien discret appliqué aux données ESPA

Estimation d'un RB discret avec les données de l'ESPA

- Ouverture d'un projet de livraison de données entre l'OFS et l'IDSIA
- L'IDSIA a utilisé un algorithme dévéloppé par eux[1]et a estimé un RB discret avec les données ESPA de 2015:
  - la structure des relations de dépendance entre 109 variables, aboutissant à un *graphe orienté acyclique (DAG)* (voir page suivante)
  - *les distributions de probabilité de toutes les variables*, qui définissent la distribution *globale* (distribution conjointe de toutes les variables)

[1]Voir M. Scanagatta, C.P. de Campos, G. Corani, M. Zaffalon, *Learning Bayesian networks with thousands of variables*, NIPS 2015 (Advances in Neural Information Processing Systems 28), 1864-1872

Graphe orienté acyclique (DAG) du réseau bayésien “ESPA 2015”



Utilisation du RB discret "ESPA 2015"

- Le RB "ESPA 2015" peut être utilisé pour répondre à des questions[2]:
  - la probabilité d'un événement précisé dans un contexte particulier $\Rightarrow$ **requête de probabilité conditionnelle**
  - la validation d'une association entre 2 variables lorsque l'influence d'une (ou plusieurs) autre(s) est annulée. Il s'agit alors d'une **requête d'indépendance conditionnelle**
  - enfin, il peut s'agir d'identifier la catégorie la plus probable d'une ou plusieurs variables. C'est alors **une recherche d'une occurrence la plus probable**, ou même de manière plus fine l'obtention de nouvelles distributions de probabilité non explicitées du RB
- Dans la suite on va montrer des exemples de requêtes de probabilités conditionnelles, en les comparant avec les estimations "classiques" (ou fréquentistes) des effectifs ($\widehat{N}$) et des fréquences conditionnelles relatives ($\widehat{f}(B0000|X)$) des statuts d'activité (B0000: actif occupé, apprenti, chômeur, non actif)

[2]Voir Denis, J.B. et M. Scutari, *Réseaux bayésiens avec R*, EDP Sciences, Collection: Pratique R, 2014

E0. Aucune condition: Population total

"n" partout plus grand que 500

- Grande région, *B023* == *"Total"*
- Sexe, *IS01* == *"Total"*
- Nationalité, *ISU1* == *"Total"*
- Cohorte de naissance (en classe), *IS02* == *"Total"*
- Degré de formation, *TBQ2* == *"Total"*

Statuts d'activité	$\widehat{N}$	$\widehat{N}_{réseau}$	$\widehat{f}(B0000\|X)$	$\widehat{P}(B0000\|X)$
Actif occupé	4'386'237	4'386'237	62.6994	62.6994
Apprenti	213'845	213'844	3.0568	3.0568
Chômeur	219'250	219'250	3.1341	3.1341
Non actif	2'176'333	2'176'334	31.1097	31.1098

E1. Trois conditions: IS01, ISU1 et TBQ2

"n" du sous-groupe Apprenti entre 10 et 89 obs.
"n" du sous-groupe Chômeur entre 90 et 489 obs.

- Grande région, *B023 == "Total"*
- **Sexe**, *IS01 == "2. Femmes"*
- **Nationalité**, *ISU1 == "2. Suisse"*
- Cohorte de naissance (en classe), *IS02 == "Total"*
- **Degré de formation**, *TBQ2 == "3. Degré Tertiaire"*

Statuts d'activité	$\widehat{N}$	$\widehat{N}_{réseau}$	$\widehat{f}(B0000\|X)$	$\widehat{P}(B0000\|X)$
Actif occupé	556'750	509'291	79.37	72.60
Apprenti	726	14'635	0.10	2.09
Chômeur	13'016	11'753	1.86	1.68
Non actif	130'991	165'804	18.67	23.64

E2. Trois conditions: IS01, ISU1 et IS02

"n" du sous-groupe Apprenti entre 0 et 4 obs.

"n" du sous-groupe Chômeur entre 90 et 489 obs.

- Grande région, *B023 == "Total"*
- **Sexe**, *IS01 == "2. Femmes"*
- **Nationalité**, *ISU1 == "2. Suisse"*
- **Cohorte de naissance (en classe)**, *IS02 == "1950 <= x < 1960"*
- Degré de formation, *TBQ2 == "Total"*

Statuts d'activité	$\widehat{N}$	$\widehat{N}_{réseau}$	$\widehat{f}(B0000\|X)$	$\widehat{P}(B0000\|X)$
Actif occupé	280'736	275'798	65.56	64.41
Apprenti	0	22	0.00	0.01
Chômeur	8'053	11'025	1.88	2.57
Non actif	139'393	141'337	32.55	33.01

E3. Quattre conditions: IS01, ISU1, IS02 et B023

"n" du sous-groupe Apprenti entre 0 et 4 obs.

"n" du sous-groupe Chômeur entre 5 et 89 obs.

- **Grande région**, *B023 == "7. Tessin"*
- **Sexe**, *IS01 == "2. Femmes"*
- **Nationalité**, *ISU1 == "2. Suisse"*
- **Cohorte de naissance (en classe)**, *IS02 == "1950 <= x < 1960"*
- Degré de formation, *TBQ2 == "Total"*

Statuts d'activité	$\widehat{N}$	$\widehat{N}_{réseau}$	$\widehat{f}(B0000\|X)$	$\widehat{P}(B0000\|X)$
Actif occupé	7'753	9'807	44.41	56.18
Apprenti	0	1	0.00	0.01
Chômeur	473	553	2.71	3.17
Non actif	9'230	7'094	52.88	40.64

E4. Quattre conditions: IS01, ISU1, IS02[2] et B023

"n" partout entre 0 et 4 obs.

- **Grande région**, *B023 == "7. Tessin"*
- **Sexe**, *IS01 == "2. Femmes"*
- **Nationalité**, *ISU1 == "2. Suisse"*
- **Cohorte de naissance (en classe)**, *"IS02 >= 2000"*
- Degré de formation, *TBQ2 == "Total"*

Statuts d'activité	$\widehat{N}$	$\widehat{N}_{réseau}$	$\widehat{f}(B0000\|X)$	$\widehat{P}(B0000\|X)$
Actif occupé	167	22	35.23	4.53
Apprenti	191	35	40.29	7.39
Chômeur	0	18	0.00	3.82
Non actif	116	400	24.47	84.26

E5. Trois conditions: IS01, ISU1 et IS02[2]

"n" partout entre 5 et 89 obs.

- Grande région, *B023 == "Total"*
- **Sexe**, *IS01 == "2. Femmes"*
- **Nationalité**, *ISU1 == "2. Suisse"*
- **Cohorte de naissance (en classe)**, *"IS02 >= 2000"*
- Degré de formation, *TBQ2 == "Total"*

Statuts d'activité	$\widehat{N}$	$\widehat{N}_{réseau}$	$\widehat{f}(B0000\|X)$	$\widehat{P}(B0000\|X)$
Actif occupé	1'337	755	11.13	6.28
Apprenti	1'669	871	13.89	7.25
Chômeur	1'121	451	9.33	3.75
Non actif	7'887	9'936	65.65	82.71

Conclusions

Conclusions

- Bilan partiel

  Les premières étapes ont été selon notre avis prometteuses, vu que:

  - l'outil est très riche en information et très léger à manipuler par rapport aux données individuelles (par exemple tout le RB présenté a un poids d'environ 500 kb)
  - jusqu'à présent on a commencé à prendre confiance avec cet outil nouveau (pour nous) et on a fait un travail de "plausibilisation"
  - il existe des potentiels usages que nous n'avons pas encore exploité

- Prochains pas

  - Réglage du RB (données ESPA des différentes années, données ESPA révisées,...)
  - Continuer avec les autres potentialités du RB (requête d'indépendence conditionnelle, d-separation, Markov-Blanket, recherche des occurences les plus probables,...)