ورشة عمل إقليمية حول استخدام بيانات التعداد لأغراض التخطيط للتنمية والبحث العلمي في البلدان العربية

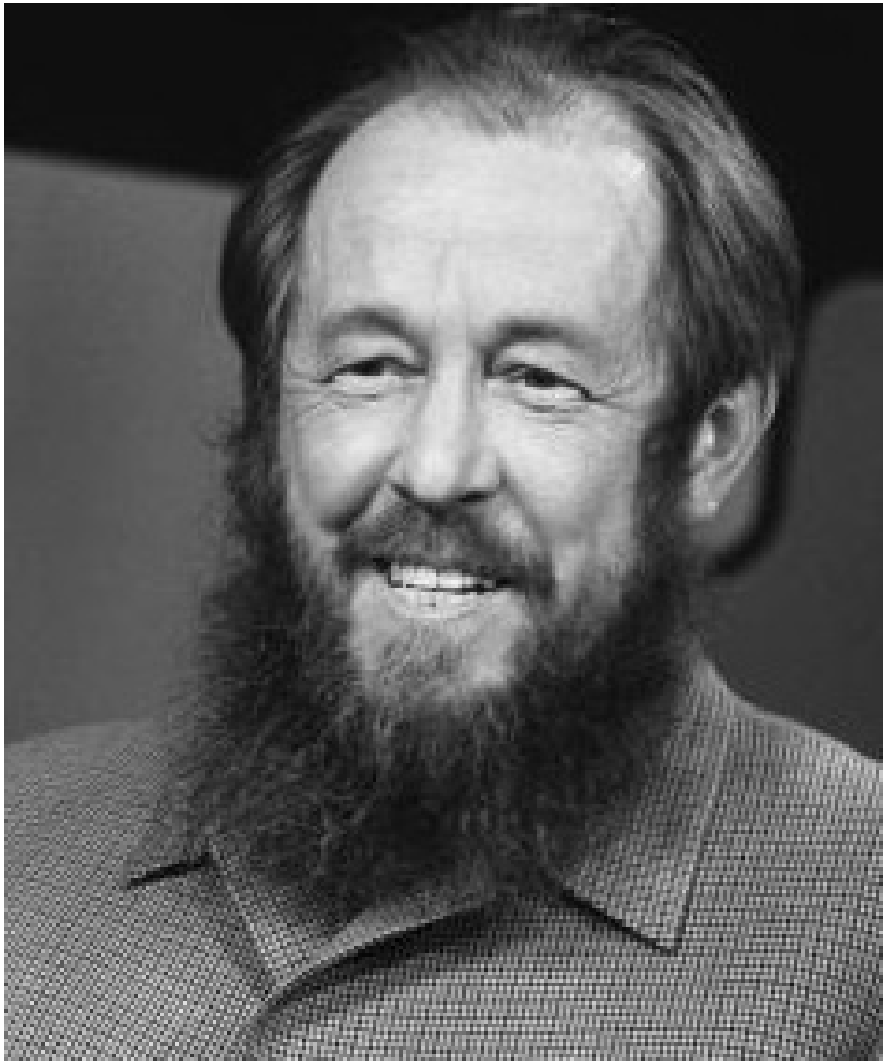
مراقبة الإفصاح في البيانات الجزئية لإحصاء 2014 من أجل نشر موسع وسري للبيانات

علي سعود

الرباط، 2 أكتوبر 2019

"حريتنا تنبني على ما يجهله الغير عن حياتنا"

ألكسندر سولجنيتسين



محتوى العرض

- تقديم
- اختيار تطبيق مراقبة الإفصاح الإحصائي
- تهييئ قاعدة الأسر وقاعدة الأشخاص
- مراقبة الإفصاح في بيانات الأسر
- مراقبة الإفصاح في بيانات الأشخاص
- الحصيلة

تقديم

1 ذكر

2 ذو بشرة فاتحة

3 ملتحي

رجل ملتحي ذو بشرة فاتحة يعاني من السرطان

⇓

إفصاح الهوية والخصائص

تقديم

- أول تجربة في المغرب لمراقبة الإفصاح الإحصائي في البيانات الجزئية (statistical disclosure control)
- الهدف: إنتاج قاعدة بيانات جزئية معالجة لإتاحتها للجميع عبر الموقع الإلكتروني
- بيانات الإحصاء مرتبة بشكل هرمي

⇐ معالجة قاعدة بيانات الأسر وقاعدة بيانات الأشخاص بالتعاقب

اختيار تطبيق مراقبة الإفصاح الإحصائي

- أهم تطبيقات مراقبة الإفصاح الإحصائي:
  - µArgus
  - sdcMicro (R)
- الاعتماد على مقارنة للتطبيقين (Templ, 2017*) تدل على أن:
  - تحميل البيانات الهرمية أسهل في sdcMicro
  - µArgus يستوجب ذاكرة وصول عشوائي (RAM) أكبر

⇐ أقل ملاءمة لمعالجة قواعد بيانات ضخمة (إحصاءات)

  - sdcMicro يمنح حرية أكبر في استخدام بعض طرق مراقبة الإفصاح، أهمها طريقة الحذف الجزئي (local suppression method)

* Templ, M. (2017). *Statistical Disclosure Control for Microdata: Methods and Applications in R*. Cham: Springer.

اختيارتطبيق مراقبة الإفصاح الإحصائي

- اقتراح البنك الدولي لمد يد المساعدة في استعمال sdcMicro تمثل أساسا في تكوين طاقم من الأطر
- وبالتالي: اختيار sdcMicro لمراقبة الإفصاح الإحصائي.

تهييئ قاعدة الأسروقاعدة الأشخاص

- التمييز بين القيم "غير المحددة" (missing values) والقيم "غير المنطبقة" (not applicable values)
- سحب عينة طبقية (stratified sample) تضم 10% من الأسر المغربية (731 ألف أسرة و 3.36 مليون نسمة) حسب الموقع الجغرافي ونوع المسكن
- تكوين مفتاح التعرف (identification key) على الأسر: الموقع الجغرافي وحجم الأسرة وفئة الأسرة ونوع الأسرة ونوع المسكن وعدد السيارات وعدد الشاحنات وعدد الجرارات

تهييئ قاعدة الأسر وقاعدة الأشخاص

- تكوين مفتاح التعرف على الأشخاص: الجنسية والنوع والعمر والحالة الزواجية وعدد الأطفال (على قيد الحياة) واللغات المحلية المستعملة ونوع النشاط والمهنة والحالة في المهنة والنشاط الاقتصادي للمشغل
- تحديد المتغيرات الحساسة (sensitive variables): عدد الأطفال المتوفين والإعاقة والحالة الزواجية وإجمالي عدد الأطفال
- سرد الترابطات (interdependencies) بين المتغيرات، كالترابط بين نوع النشاط والمهنة والحالة المهنية والنشاط الاقتصادي للمشغل ومكان العمل ووسيلة النقل إلى مكان العمل

⇐ تعديل المتغيرات المرتبطة في آن واحد في حالة ما إذا تغير أحدهم

مراقبة الإفصاح في بيانات الأسر

المستويات الجغرافية الرئيسية بالمغرب:

1. المستوى الوطني
2. مستوى الجهات (12 وحدة)
   - وسيط عدد سكان الجهات: 2.6 مليون نسمة (7.7% من المجموع)
3. مستوى العمالات والأقاليم (75 وحدة)
   - وسيط عدد سكان العمالات والأقاليم: 370 ألف نسمة (1.1% من المجموع)
4. مستوى الجماعات والمقاطعات (1,538 وحدة)
   - وسيط عدد سكان الجماعات والمقاطعات: 10 آلاف نسمة (0.03% من المجموع)

مراقبة الإفصاح في بيانات الأسر

3 مراحل كبرى لمراقبة الإفصاح:

1. قياس خطر الإفصاح (disclosure risk) الأولي
2. تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي
3. تقييم مدى تغيير البيانات إثر مراقبة الإفصاح

مراقبة الإفصاح في بيانات الأسر

قياس خطر الإفصاح الأولي

اختيار مجموعة من مقاييس الخطر:

- عدد الأسر غير الخاضعة لقاعدة المجهول من الدرجة الخامسة (k-anonymity ; k=5)
  $A_5(init) = 47\ 504\ (6.5\%)$
- الخطر العام في قاعدة البيانات $R(init) = 1,0\%$
- أقصى خطر وحدوي $r_{max}(init) = 25,6\%$
- عدد الأسر في خطر عند عتبة 5% $U_{5\%}(init) = 29\ 026\ (4,0\%)$
- عدد الأسر في خطر عند عتبة 20% $U_{20\%}(init) = 16\ 980\ (2,3\%)$

مراقبة الإفصاح في بيانات الأسر

تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي

- تطبيق عتبة للنشر قدرها 1‰ من السكان (حوالي 30 ألف نسمة)

  ⇐ تجميع كل إقليم يقطن به أقل من 30 ألف شخص مع أصغر إقليم مجاور له وينتمي إلى نفس الجهة

- إزالة الأسر الرحالة والأسر المكونة من أشخاص بدون مأوى (0.8‰) وكذا الأسر الكبيرة –المكونة من 20 شخص أو أكثر– (0.9‰)

- إعادة ترميز شاملة (global recoding) لبعض متغيرات مفتاح التعرف، حسب تجانس اختيارات الجواب

مراقبة الإفصاح في بيانات الأسر

تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي

- تطبيق طريقة الحذف الجزئي (local suppression) على مفتاح التعرف مع ترتيب متغيراته حسب قابليتهم للحذف، بهدف إلى أن تصبح قاعدة البيانات مجهولة من الدرجة الخامسة (5-anonymous)
- حذف معلومات إضافية من المتغيرات المترابطة مع متغيرات مفتاح التعرف التي وقع بها حذف
- إزالة الأسر التي حذفت بها المعلومة حول حجم الأسرة (42 أسرة)
- تكرير عملية الحذف الجزئي 4 مرات متعاقبة للحصول على قاعدة بيانات مجهولة من الدرجة الخامسة ⇐ تتطلب وقتا مهما (ساعات عديدة)

مراقبة الإفصاح في بيانات الأسر

تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي

$A_5(init) = 47\,504\ (6.5\%)$	$A_5(fin) = 0$
$R(init) = 1.0\%$	$R(fin) = 0.2\%$ (↘81%)
$r_{max}(init) = 25.6\%$	$r_{max}(fin) = 2.4\%$ (↘91%)
$U_{5\%}(init) = 29\,026\ (4.0\%)$	$U_{5\%}(fin) = 0$
$U_{20\%}(init) = 16\,980\ (2.3\%)$	$U_{20\%}(fin) = 0$

مراقبة الإفصاح في بيانات الأسر

تقييم مدى تغيير البيانات إثر مراقبة الإفصاح

- حساب عدد القيم المحذوفة
  - ⇐ المتغير الذي عرف أكبر عدد هو نوع المسكن: 2.1% من القيم
  - ⇐ متوسط عدد القيم المحذوفة بالمتغيرات المعنية: 0.7% من القيم
- مقارنة الأرقام الرئيسية للإحصاء (41 مؤشر) قبل وبعد مراقبة الإفصاح، على الصعيد الوطني والجهوي والإقليمي إضافة إلى وسط الإقامة، وذلك بحساب القيم المطلقة للفوارق
  - ⇐ المتوسط: بين 0.08% (وطني حضري) و 0.70% (إقليمي قروي)
  - ⇐ الانحراف المعياري: بين 0.16% (وطني) و 1.86% (إقليمي قروي)

مراقبة الإفصاح في بيانات الأشخاص

قياس خطر الإفصاح الأولي

اختيار نفس مقاييس الخطر مع أخذ البنية الهرمية للمعطيات بعين الاعتبار:

- عدد الأشخاص غير الخاضعين لقاعدة المجهول من الدرجة الخامسة

$A_5(init) = 235\ 898\ (7.0\%)$

- الخطر العام في قاعدة البيانات

$R(init) = 5.1\%$

- أقصى خطر وحدوي

$r_{max}(init) = 96.1\%$

- عدد الأشخاص في خطر عند عتبة 5%

$U_{5\%}(init) = 748\ 026\ (22.3\%)$

- عدد الأشخاص في خطر عند عتبة 20%

$U_{20\%}(init) = 339\ 696\ (10.1\%)$

مراقبة الإفصاح في بيانات الأشخاص

تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي

- إعادة ترميز شاملة (global recoding) لبعض متغيرات مفتاح التعرف، حسب تجانس اختيارات الجواب
- حماية الطابع الحساس للأمهات العازبات بتطبيق طريقة التوزيع العشوائي البعدي (PRAM: Post-randomization method) على الحالة الزواجية

⇐ تغيير 0.4‰ من قيم الحالة الزواجية اعتمادا على مصفوفة عشوائية (stochastic matrix) تتضمن احتمالات تغيير اختيارات الجواب

مراقبة الإفصاح في بيانات الأشخاص

تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي

- تطبيق طريقة الحذف الجزئي (local suppression) على مفتاح التعرف مع ترتيب متغيراته حسب قابليتهم للحذف، بهدف إلى أن تصبح قاعدة البيانات مجهولة من الدرجة الخامسة (5-anonymous)
- حذف معلومات إضافية من المتغيرات المترابطة مع متغيرات مفتاح التعرف التي وقع بها حذف

مراقبة الإفصاح في بيانات الأشخاص

تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي

A_5(init) = 235 898 (7.0%)

R(init) = 5.1%

r_{max}(init) = 96.1%

$U_{5\%}$(init) = 748 026 (22.3%)

$U_{20\%}$(init) = 339 696 (10.1%)

A_5(fin) = 0

R(fin) = 0.4% (**↘92%**)

r_{max}(fin) = 16.6% (**↘83%**)

$U_{5\%}$(fin) = 7 357 (0.2%) (**↘99%**)

$U_{20\%}$(fin) = 0

مراقبة الإفصاح في بيانات الأشخاص

تطبيق طرق مراقبة الإفصاح بغية تقليص خطر الإفصاح الأولي

تقييم التنوع من الدرجة الثانية (l-diversity ; l=2) بالنسبة للمتغيرات الحساسة:

- 32 شخصا (10^{-5}%) معنيون بإفصاح عدد أطفالهم المتوفين ولا أحد معني بإفصاح حالة إعاقته

⇐ حذف المعلومات حول عدد الأطفال المتوفين بالنسبة للأشخاص المعنيين

⇐ الحصول على قاعدة بيانات متنوعة من الدرجة الثانية (2-diverse) فيما يخص المتغيرات الحساسة

مراقبة الإفصاح في بيانات الأشخاص

تقييم مدى تغيير البيانات إثر مراقبة الإفصاح

- حساب عدد القيم المحذوفة
  - ⇐ المتغير الذي عرف أكبر عدد هو اللغات المحلية المستعملة: 0.9% من القيم
  - ⇐ متوسط عدد القيم المحذوفة بالمتغيرات المعنية: 0.1% من القيم
- مقارنة الأرقام الرئيسية للإحصاء (85 مؤشر) قبل وبعد مراقبة الإفصاح، على الصعيد الوطني والجهوي والإقليمي إضافة إلى وسط الإقامة والنوع، وذلك بحساب القيم المطلقة للفوارق
  - ⇐ المتوسط: بين 0.06% (وطني ذكور) و 0.65% (إقليمي قروي)
  - ⇐ الانحراف المعياري: بين 0.11% (وطني ذكور) و 3.33% (إقليمي قروي)

الحصيلة

- حماية دقيقة للسكان

⇐ كسب ثقة المستجوبين وتشجيع فتح البيانات الجزئية

- خلال مايو 2019، نشر عينة من البيانات الجزئية لإحصاء 2014 دون المس بخصوصية الأشخاص، وذلك للعموم عبر الموقع الإلكتروني:

https://www.hcp.ma/downloads/RGPH-2014-Microdonnees-anonymisees-Open-Data_t21400.html

- قاعدتي بيانات الأسر والأشخاص بقطع نص (text format) وبقطع SPSS وبقطع STATA
- البيانات الوصفية (metadata)
- الاستمارة (questionnaire)
- وثيقة منهجية لمراقبة الإفصاح الإحصائي (methodology note on SDC)

علي سعود

a.saoud@hcp.ma