Le temps dans les analyses quantitatives de données

Présentation pour le cours SOL6210, Analyse quantitative avancée

© Claire Durand, 2022

Le temps ... en soi

- Tous les processus, accès à la jeunesse comme à l'âge adulte, modification des fonctions assumées, des perceptions, etc., évolution de la criminalité, du chômage ou de la confiance, se déroulent dans le temps.
- Il est donc primordial et fort pertinent d'en tenir compte dans les analyses
- Principe fondamental: Le changement s'explique par le changement

Définir, recueillir, décider (1)

Les décisions à prendre

- Lorsque l'on travaille avec le temps, il faut
  - Décider du "groupe à risque", et par conséquent, de la "période à risque".
  - Donc, non seulement auprès de qui on recueille les données mais à partir de quand et jusqu'à quand ou... portant sur quelle période (données rétrospectives)

Définir, recueillir, décider (2)

Les types de données

- Plusieurs types de données:
  - Données de type panel: les mêmes personnes interrogées à plusieurs reprises
    - Sur leur situation avant l'entrevue
    - Sur leur situation au moment de l'entrevue
    - Sur leur situation entre les moments d'entrevue
  - Données de type longitudinal comprenant plusieurs échantillons
    - Qui peuvent avoir été recueillis indépendamment (plusieurs sondages auprès d'échantillons différents mais certaines questions identiques)
    - Qui peuvent être recueillies sous forme de sondage roulant (rolling cross-section): on met sur le terrain un nouveau sous-échantillon à chaque jour (pendant une campagne électorale par exemple).

Définir, recueillir, décider (2)

Les types de données (suite)

- Données de type archive
  - Statistiques institutionnelles,
  - Données économiques, taux de chômage, évolution du PIB, des salaires, etc.
  - Données sur les taux de criminalité, de mortalité, etc.

Les problèmes associés

Deux problèmes principaux

- Dépendance des données dans le temps, autocorrélation
  - Ce qui fait que le postulat de base des analyses de régression ordinaire n'est pas respecté à sa face même. Ceci biaise l'estimation de la variance.
- Données manquantes à certains moments, censure et troncature
  - Personnes qui "disparaissent", absence d'informations pour certaines unités à certains moments: il faut connaître les raisons de l'absence de données et s'assurer que cette absence est aléatoire, que ça ne biaise pas les analyses

Que faire? (1)

D'abord et avant tout décrire...Visualiser aide à se représenter la situation. A cette étape, on peut repérer des problèmes

- Graphique des transitions (...) (DeGenne, LeBourdais, Renaud, etc. pour exemples)
- Tapis (ex: Degenne (site du CIQSS): description de l'évolution de la situation des finissants)
- Tables de survie (Renaud, Durand, etc.)
- Graphiques de séries chronologiques (Durand, Larochelle et Blais, 2005; Durand, 2008, 2011)
- Régressions locales (Durand, à partir de 2014).
- Classifications de trajectoires (Durand et Lacourse; Durand, Pelletier, Wutchiett)

Activité post d.u.t (2 ans) De Genne (2003)

Rose = CDD; Rouge = CDI; Violet = alternance; Bleu = intérim;
Orange = stage; Jaune = Service national;Noir = chômage; Blanc = autre situation;
Vert = études ; Cyan = non réponse ou inactivité

Que faire (2)

Quelques décisions importantes

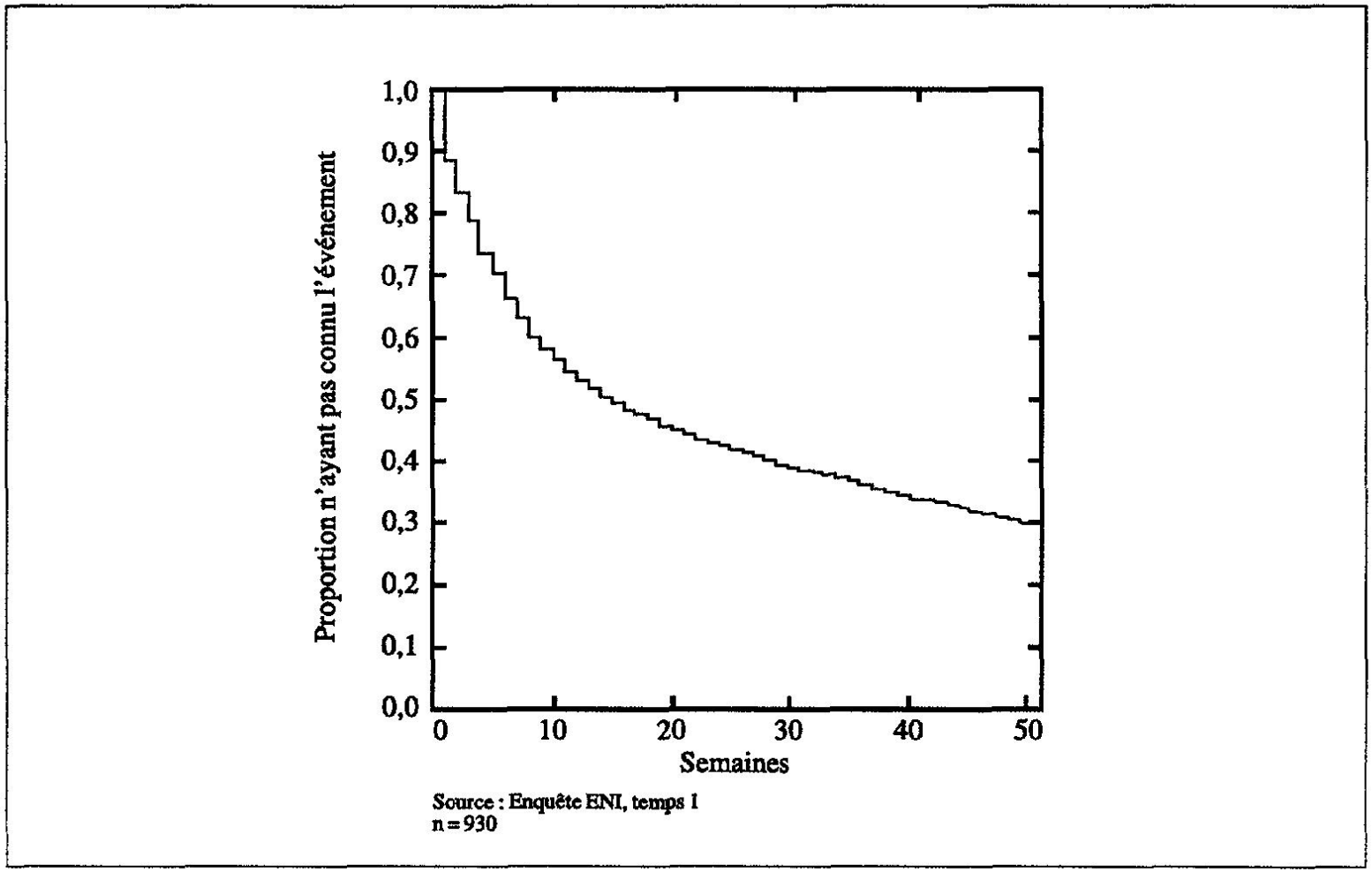
- Quel est, quels sont, les événements d'intérêt?
- Qu'est-ce qui donne la mesure du temps? (jours, mois, années, essais)
- Le temps est-il discret ou continu?
- Les événements qui prédisent ou expliquent la variable dépendante se modifient-ils dans le temps?
- Quelle est la forme de l'évolution dans le temps?

Que faire? (3)

Multiples analyses disponibles selon la situation

- Tables de survie et régressions de survie
  - S'utilisent quand on a des informations sur un événement d'intérêt qui survient à un moment donné; on s'intéresse à ce qui explique la ***rapidité*** de transition à l'état d'intérêt.
  - Exemple:
    - Qu'est-ce qui explique la rapidité avec laquelle un immigrant se trouve un travail en arrivant dans un pays? A peu près tous les immigrants finissent par se trouver un travail (d'où peu d'intérêt à la régression logistique); ce qui nous intéresse donc, c'est la ***rapidité d'accession***. On peut aussi se poser la question de la rapidité d'accession à un travail de tel type, dans telle langue, etc.

Table de survie Renaud (1992)



Graphique 1 Table des entrées dans un premier emploi au Québec

Régression de survie, prédicteurs rapidité d'accès à un emploi

Renaud (1992)
(voir suite p. Suivante)

Tableau 2
L'accès au premier emploi : les variables acquises avant la migration

Modèle	1	2	3
Log. de vraisemblance	– 3865.6	– 3822.5	– 3819.9
Âge (années)	**– 0.037 ***	**– 0.040 ***	**– 0.040 ***
Sexe (1 = homme)	0.541 *	**0.589 ***	**0.588 ***
Scolarité (années)	0.000	0.000	0.000
Immigrant «famille»	**0.354 ***	**0.301**	0.297
Immigrant «indépendant»	**0.479 ***	**0.424 ***	**0.420 ***
Expérience de travail pré QC	0.350 *	**0.284 ***	**0.282 ***
Connaissance à l'arrivée			
Connaissance de l'anglais	– 0.014	– 0.002	0.000
Connaissance du français	– 0.025	– 0.115	– 0.104
Cours en cours			
enseig. régulier t. plein		**– 0.714 ***	**– 0.695 ***
COFI t. plein		**– 1.700 ***	**– 1.681 ***
profes. et al t. plein		**– 1.972 ***	**– 1.965**
français t. partiel		0.401	0.403
anglais t. partiel		0.288	0.289
COFI t. partiel		– 0.252	– 0.249
profes. et al t. partiel		0.309	0.325
Diplomation			
enseig. régulier t. plein		0.259	—
COFI t. plein		**1.501 ***	**0.834 ***
profes. et al t. plein		0.707	—
français t. partiel		0.903	0.899
anglais t. partiel		0.490	0.453
COFI t. partiel		– 0.831	– 0.707
profes. et al t. partiel		0.679 *	—

Suite du tableau précédent (3ème colonne)

Diplomation selon la langue du cours	
régulier anglais	**1.469 ***
profes. et al t. pl. anglais	0.608
profes. et al t. pa. anglais	0.069
régulier français	– 0.495
profes. et al t. pl. français	0.826
profes. et al t. pa. français	**0.957 ***

* 0.05 (n=930)

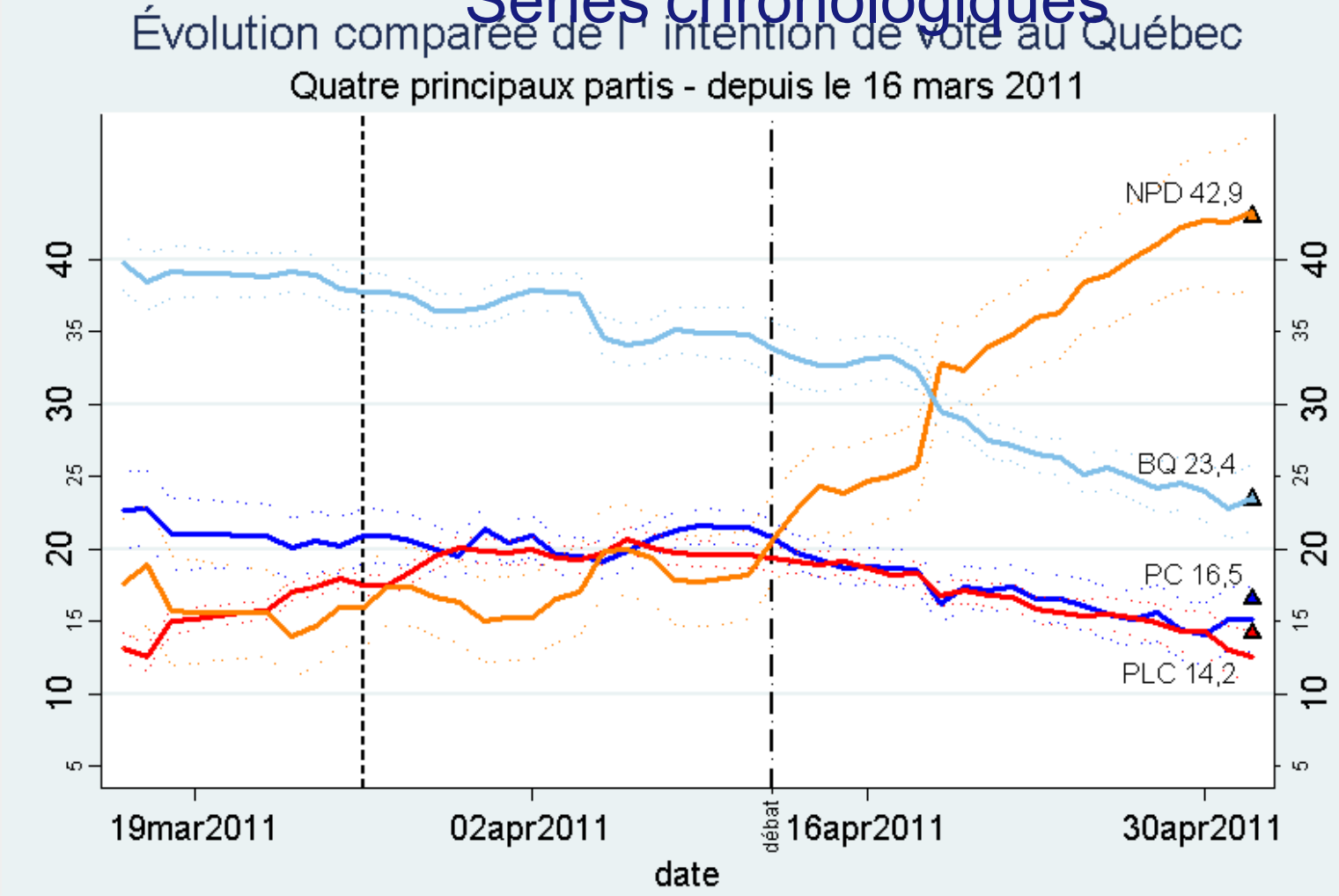
Que faire? (4)

Multiples analyses disponibles selon la situation

- Séries chronologiques simples ou croisées
  - Les données sont des informations habituellement agrégées **pour tous les moments de mesure**. S'il manque des données à un moment, il y a des procédures pour "intrapoler".
  - On peut voir si une série est influencée par des événements qui surviennent ou par d'autres séries d'événements, si les séries se distinguent d'une unité -- pays, etc.-- à une autre.
  - Exemple: suite des taux de chômage pour chaque mois, suite des sondages pendant une campagne électorale, suite de taux de chômage et de taux d'inflation (relation entre les deux?), etc.

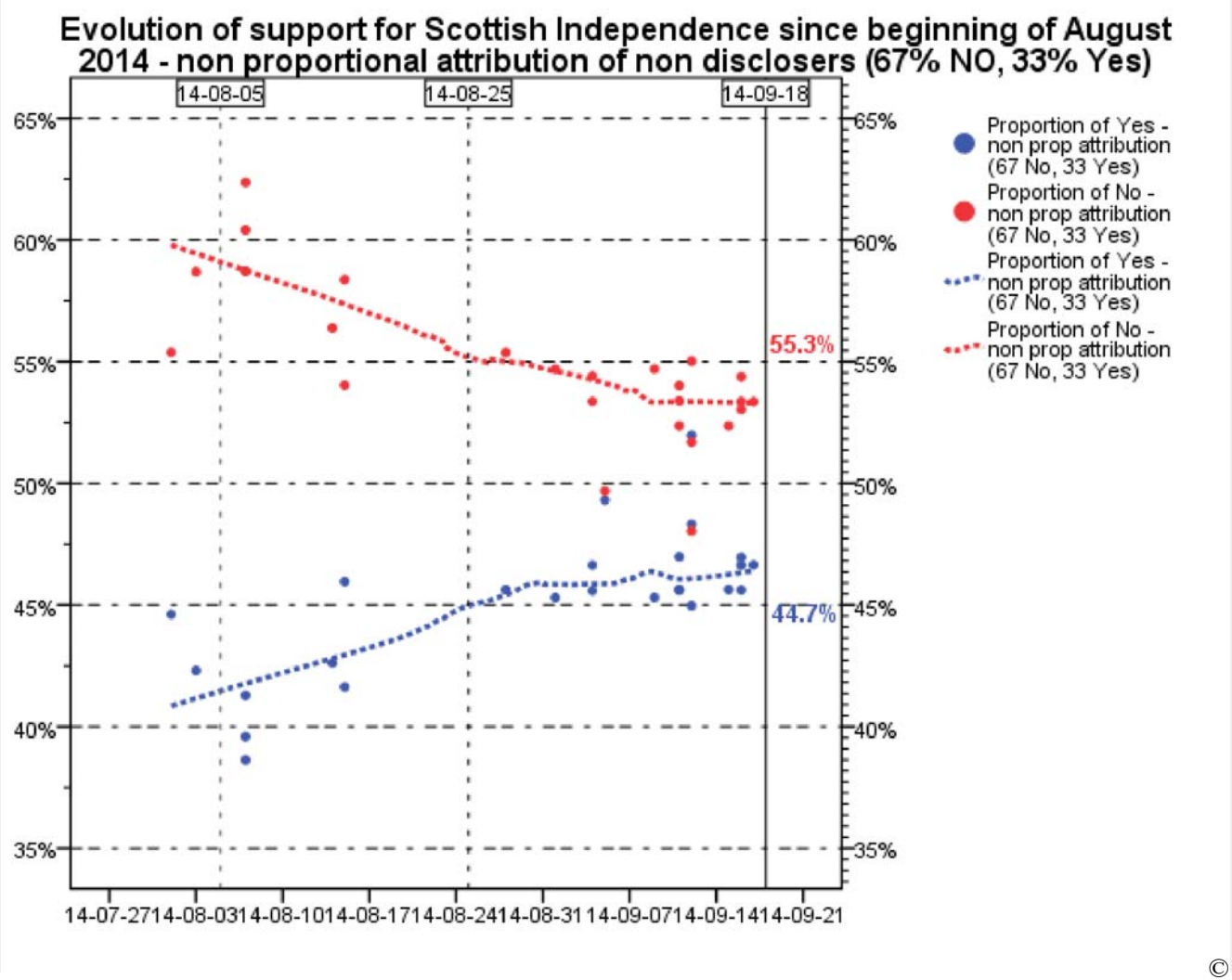
Évolution de l'intention de vote (Canada 2011 au Québec), sondages publiés

Séries chronologiques

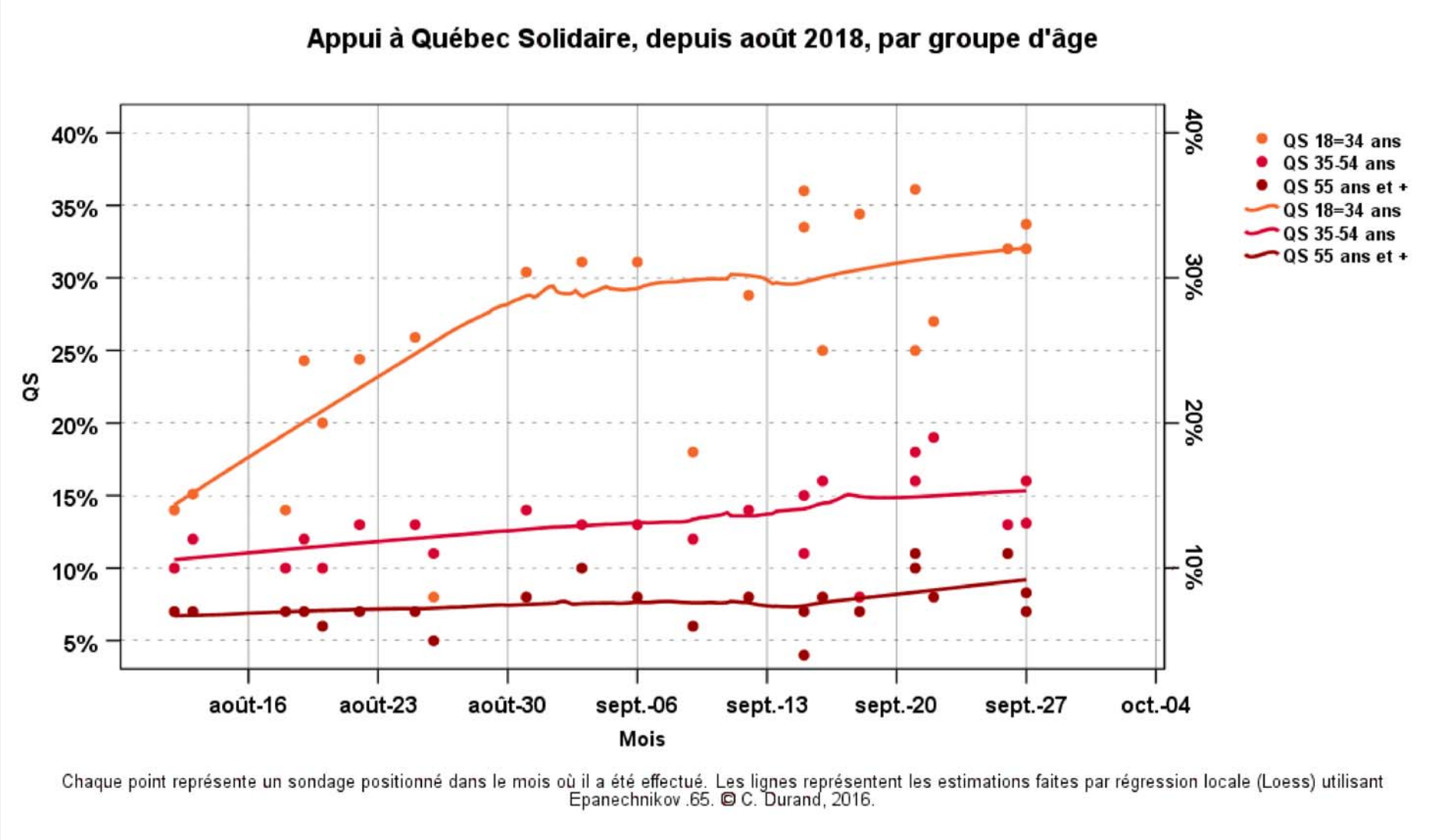


Évolution de l'intention de vote Référendum Écosse 2014

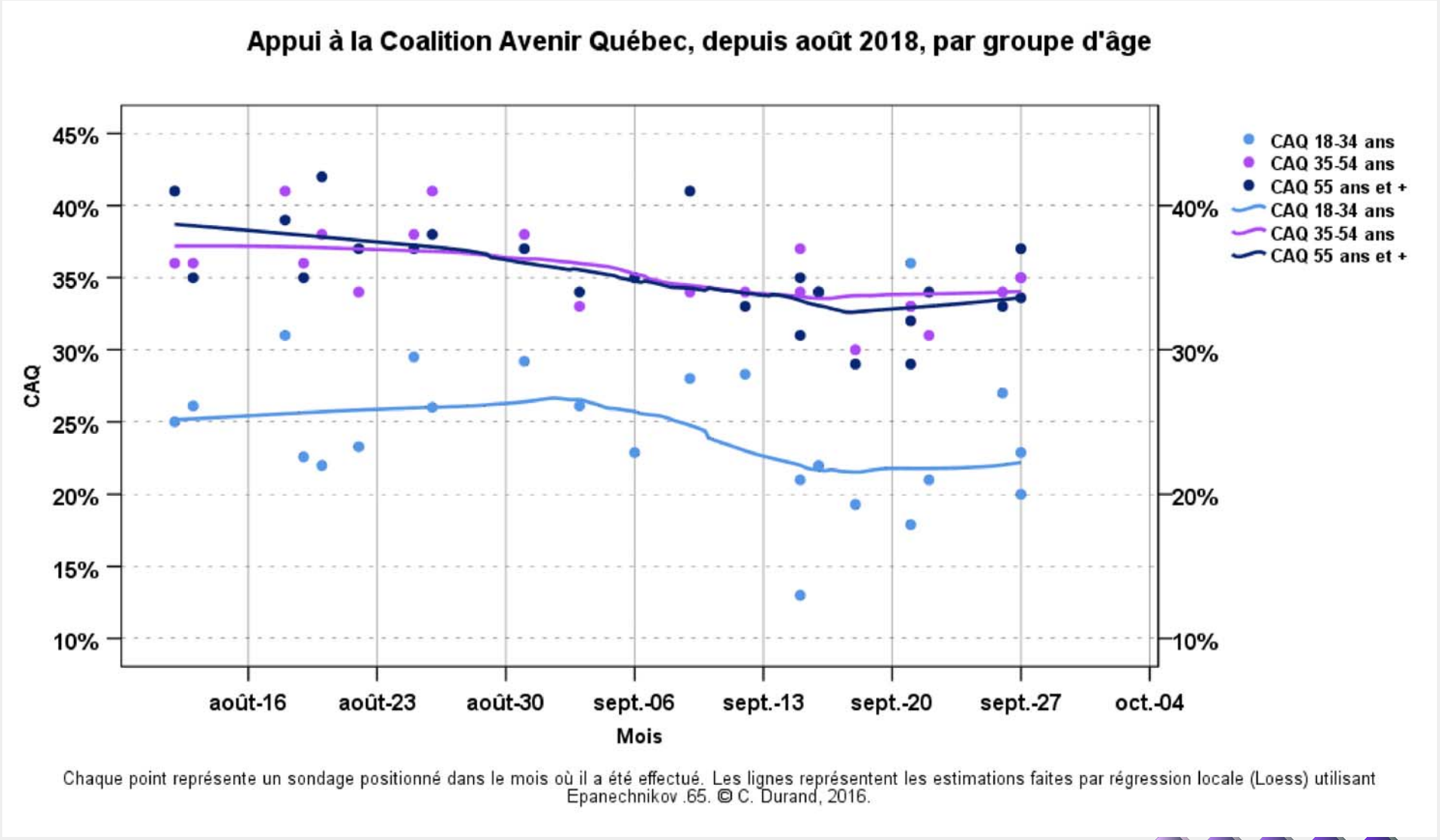
après répartition non proportionnelle des discrets



Évolution de l'intention de vote pour Québec Solidaire, selon l'âge

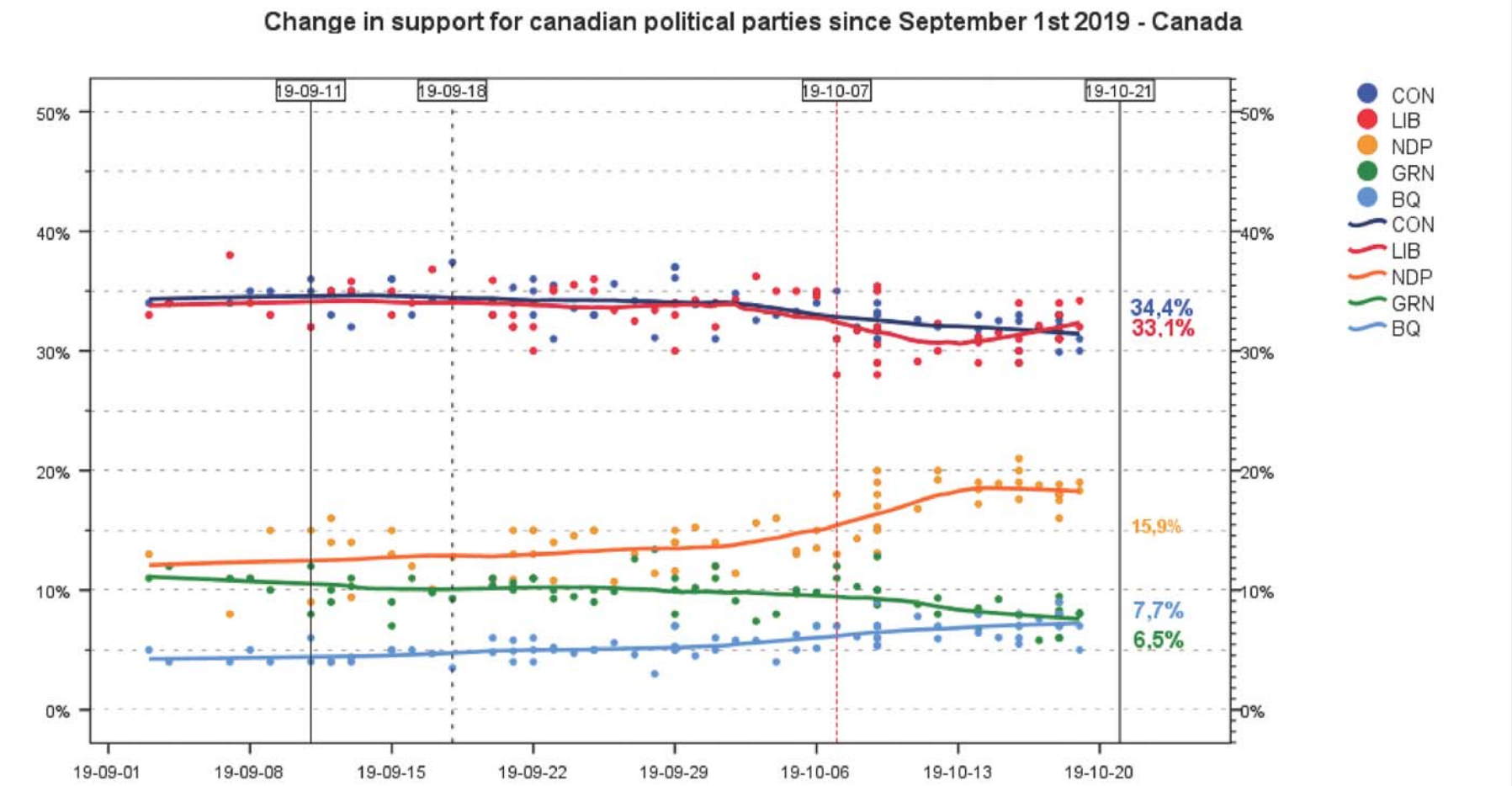


Évolution de l'intention de vote pour la Coalition Avenir Québec, selon l'âge

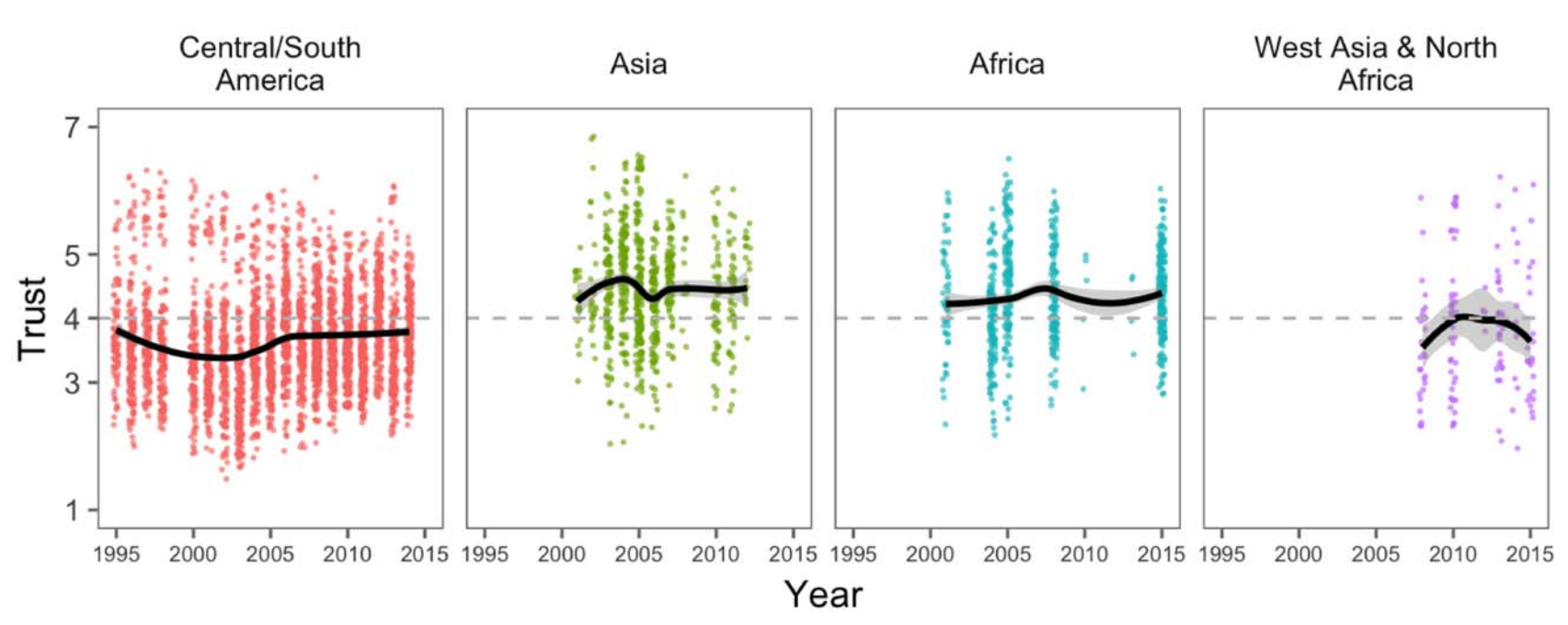


Évolution de l'intention de vote - Canada 2019 - selon les sondages publiés

Régressions locales

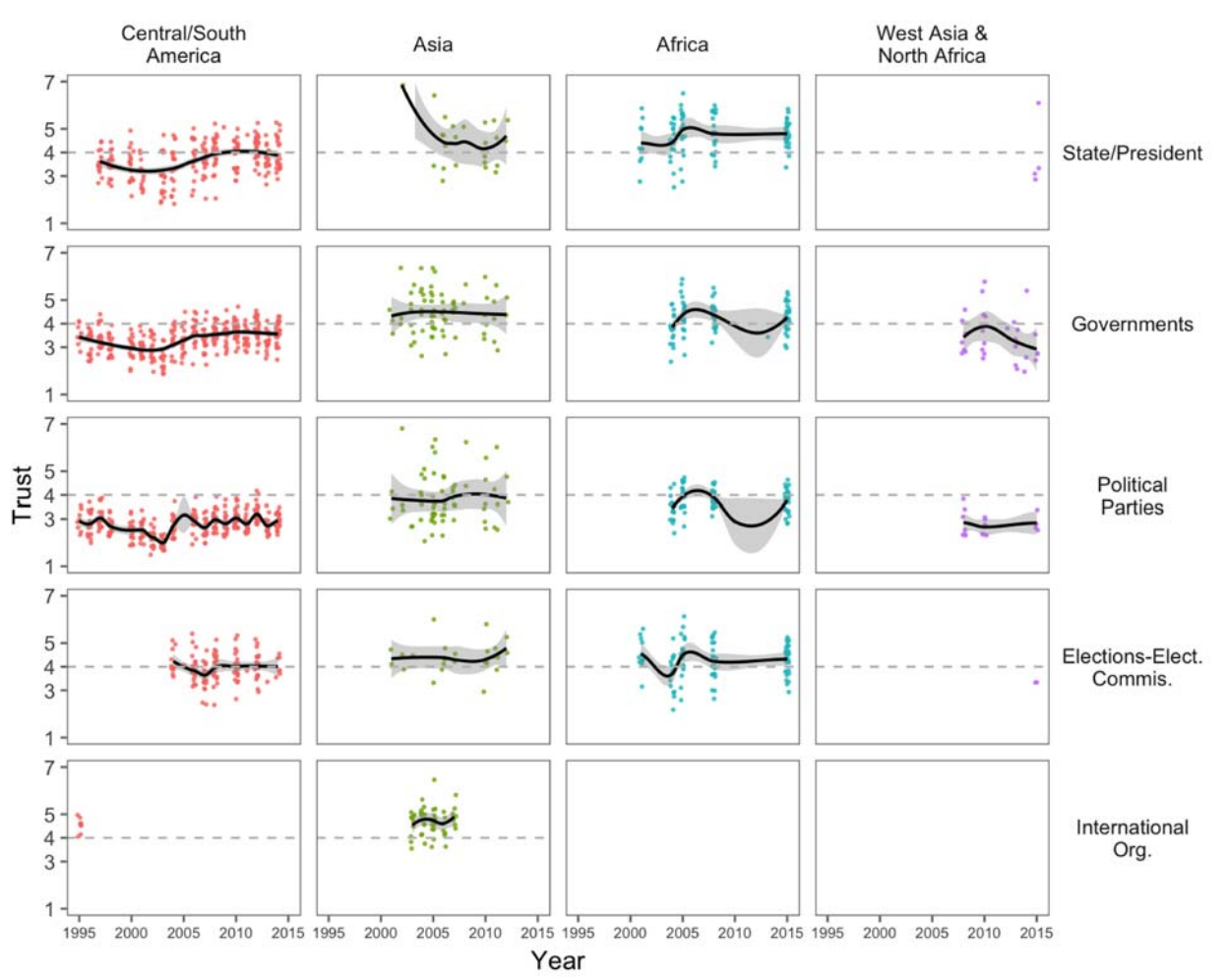


Une vue synthétique de l'évolution de la confiance institutionnelle par région du monde.



- En moyenne, la confiance est stable
- Plus élevée en Asie et en Afrique qu'en Amérique latine.
- A diminué beaucoup depuis 2011 en Afrique du Nord & Asie de l'Ouest.

Confiance dans les institutions politiques.



- Trust lower in South/Central America & WANA.
- Political parties, lowest in South/ Central America & WANA.
- Drop in trust in gvt in WANA, in State/ President in Asia.

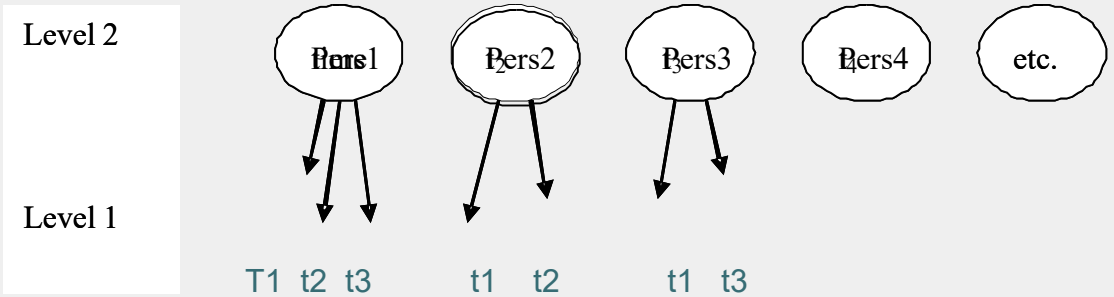
Que faire? (5)

Multiples analyses disponibles selon la situation

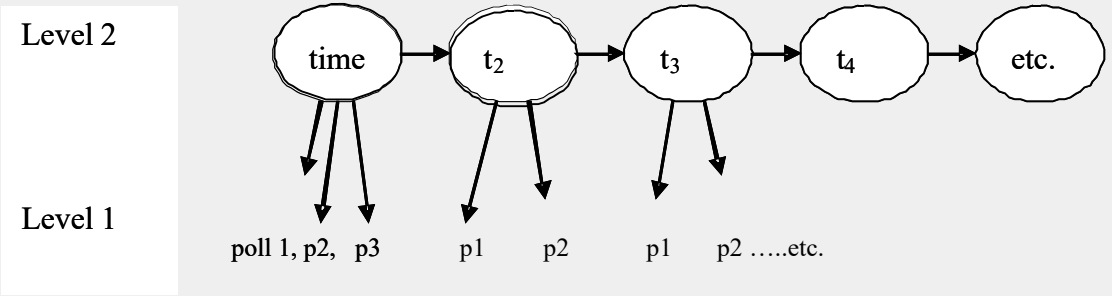
- Analyses multi-niveaux longitudinales
  - Dans ce cas, le temps est considéré comme un niveau: les diverses mesures prises sont “nichées” dans les individus qui peuvent eux-mêmes être nichés dans des unités (classes, équipes de travail, familles,...).
  - L’intérêt est la flexibilité de la méthode, entre autres parce qu’il n’est pas obligatoire d’avoir des mesures à chaque moment et au même moment pour tous les sujets.
  - Il faut que le niveau supérieur (2 ou 3) soit un échantillon (n> 40).
  - Exemple: évolution de l’emploi durant un certain temps, évolution de la confiance institutionnelle dans le temps pour un certain nombre de pays (échantillon de pays ou de périodes).

Modèle multiniveaux longitudinal

- Le temps peut être au niveau 1, par exemple, personnes (niveau 2) avec mesures prises à divers moments (niveau 1)



Le temps peut être au niveau 2, par exemple, mois (niveau 2) avec résultats de sondages faits à chaque mois (niveau 1)



Évolution de la confiance institutionnelle

Trust in institutions - basic models										
	Model 0		**Model 1**		**Model 2**		**Model 3**		**Model 4**	
Intercept	4.12725	***	4.527	***	4.290	***	4.452	***	3.817	***
Level Measure										
Media (REF)										
Church			0.678	***	0.678	***	0.678	***	0.678	***
Trade Unions			-0.785	***	-0.785	***	-0.785	***	-0.785	***
ONG- Civil Society			-0.449	***	-0.448	***	-0.449	***	-0.449	***
Army-police			-0.265	***	-0.266	***	-0.266	***	-0.266	***
Public Admin.			-0.504	***	-0.503	***	-0.504	***	-0.504	***
Judiciary			-0.481	***	-0.481	***	-0.481	***	-0.481	***
Finance			-0.326	***	-0.328	***	-0.326	***	-0.326	***
Enterprises			-0.454	***	-0.454	***	-0.454	***	-0.454	***
State/President			-0.219	***	-0.220	***	-0.220	***	-0.220	***
Governments			-0.606	***	-0.606	***	-0.606	***	-0.606	***
Political Parties			-1.151	***	-1.151	***	-1.151	***	-1.151	***
Elections- Elect. Commis.			-0.378	***	-0.377	***	-0.378	***	-0.378	
International Org.			-0.198	***	-0.196	***	-0.198	***	-0.198	***
Level Respondent										
woman					0.004	ns	0.004	ns	0.004	ns
Young (Less than 30)					0.009	**	0.009	**	0.009	**
time										
Old (60 plus)					0.095	***	0.094	***	0.095	***
time										
Prop_Non-resp.					0.003	***	0.004	**	0.003	**
Level Country-Year										
Time							0.001	ns	0.001	ns
Time2							0.001	*	0.001	*
Level Country-Source										
Central/South America (REF)										
Asia									1.022	***
Africa									0.875	***
West Asia N. Africa									0.496	**
Answer Scale (7 pts)										
									0.425	**
Variance										
Measures	2.553	62.9%	2.380	60.6%	2.380	60.7%	2.380	60.7%	2.380	63.1%
Respondents	1.097	27.0%	1.112	28.3%	1.109	28.3%	1.109	28.3%	1.109	29.4%
Country-Year	0.106	2.6%	0.108	2.8%	0.109	2.8%	0.107	2.7%	0.109	2.9%
Country-Source	0.303	7.5%	0.325	8.3%	0.324	8.3%	0.327	8.3%	0.176	4.7%
Total	4.059		***3.925***		***3.922***		***3.924***		***3.775***	
			6.8%		0.2%		1.3%		46.3%	

Focus sur les niveaux 2 & 3

Trust in institutions - basic models				
	Model 2		**Model 3**	
Intercept	4,290	***	4,452	***
Level Respondent				
woman	0,004	ns	0,004	ns
Young (Less than 30)	0,009	**	0,009	**
time				
Old (60 plus)	0,095	***	0,094	***
time				
Prop_Non-resp.	0,003	***	0,004	**
Level Country-Year				
Time			0,001	ns
Time2			0,001	*
Variance				
Measures	2,380	60,7%	2,380	60,7%
Respondents	1,109	28,3%	1,109	28,3%
Country-Year	0,109	2,8%	0,107	2,7%
Country-Source	0,324	8,3%	0,327	8,3%
Total	***3,922***		***3,924***	
	0,2%		1,3%	

- Individual level:
  - Sex is not significant
  - *Compared to middle age:*
    - being less than 30: +.009;
    - being 60+: +.094 .
  - Prop. Non-response: +.003.
- **Niveau pays-année:**
  - **Le temps au carré est significatif.**
- Variance explained: minimal

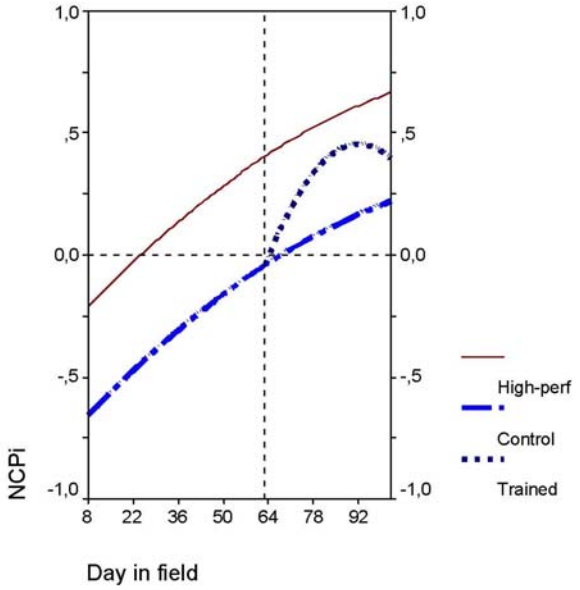
Que faire? (6)

Multiples analyses disponibles selon la situation

- Analyse des trajectoires
  - Il s'agit de faire une classification des trajectoires "individuelles" -- ca peut être des individus mais aussi des groupes, des pays, etc. -- pour en arriver à des regroupements de parcours.
  - Méthode en développement, relativement récente mais en voie d'être intégrée dans les principaux logiciels.
  - Exemple: Les trajectoires de délinquance entre l'âge de 5 ans et de 18 ans, au moyen de mesures similaires prises à divers moments durant cette période.
  - Problème: prédire le passé avec le futur.

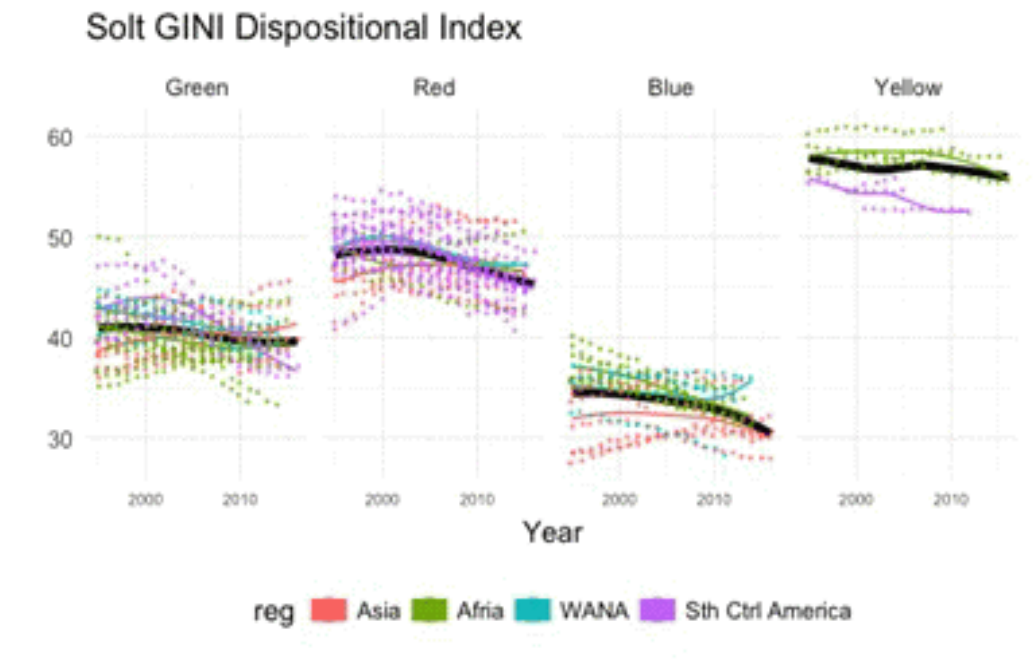
Analyses de trajectoires, avant et après formation, avec groupe contrôle (bleu)

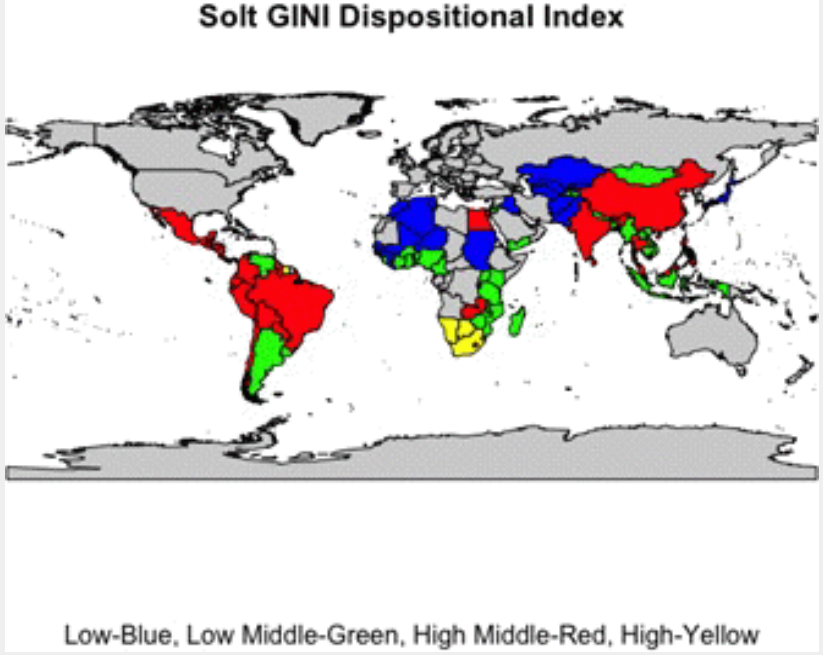
Trajectoires de performance (NCPI) des interviewers.
Groupe rouge: bonne performance, non formés
Groupe bleu: faible performance, non formés
Groupe gris: faible performance, formés



Analyse de trajectoires de mesures

L'évolution de certaines mesures dans le temps: http://www.mapageweb.umontreal.ca/durandc/Recherche/Publications/confiance/WAPOR2018_CD.pdf





Que faire? (7)

Multiples analyses disponibles selon la situation

- Analyse de variance pour mesures répétées
  - Analyse relativement traditionnelle en psychologie. D'une certaine manière, c'est la base.
  - L'idée est de faire une intervention et de mesurer la variable dépendante à divers moments fixés pour analyser l'évolution entre les moments.
  - Exemple :
    - Mesure de la dépression à divers moments durant un processus thérapeutique, évolution de la performance en mathématique avant et après une intervention de remédiation, etc.
  - Problème: On ne peut garder que les cas pour lesquels on a de l'information à **tous** les temps de mesure.

Quelle forme prend l'évolution dans le temps?

- Pour la plupart des analyses, il est très important de se demander quelle forme prend l'évolution dans le temps en soi ou suite à un événement (voir Effet "Bouchard" pendant la campagne référendaire de 1995 au Québec, Durand, 2008)
- L'évolution peut être simplement linéaire mais elle peut aussi être quadratique en "U", cubique (en forme de dos de poisson), voir Applied Longitudinal Data Analysis, Singer & Willett, 2003.
- Un événement peut être de provoquer un saut. Il peut aussi provoquer une modification de l'évolution: accélération, plafonnement, etc.

Les types d'analyses de prédiction

Types de variables et modes d'entrée

- Pour toutes ces procédures, plus ou moins facilement selon les logiciels utilisés, les variables indépendantes et dépendantes peuvent être de différents types.
- Les variables indépendantes peuvent être fixes ou varier dans le temps.
- Pour toutes ces procédures, les variables indépendantes peuvent être entrées ensemble (régression standard) ou de façon hiérarchique/ séquencielle.

Avantages et inconvénients

Quel type d'analyse choisir?

- Le choix de l'analyse dépend de plusieurs facteurs, dont la question de recherche, le type de données, les finalités de l'analyse.
- La plupart du temps, le choix du type d'analyse s'impose étant donné les données et la question de recherche. Dans certains cas, plusieurs analyses sont possibles. Des informations différentes seront mises en évidence selon le type d'analyse mais les conclusions statistiques seront rarement différentes.

Avantages et inconvénients

Interventions sur les données

- Dans les analyses de ce type, il est souvent nécessaire de faire des interventions sur les fichiers
  - Pour les analyses de survie (tables, régressions de survie), il faut “rectangulariser” le fichier (voir procédure restructurer de SPSS) (voir travail sur les fichiers http://www.mapageweb.umontreal.ca/durandc/menuMethodesQuantitatives.html#travail_fichier).
  - Il faut parfois créer des variables qui indiquent le moment où un événement survient.
  - Pour les analyses multi-niveaux avec HLM, il faut faire un fichier par niveau (procédures Agréger ou Restructurer dans SPSS)

Choix des logiciels

- Le logiciel STATA est probablement le plus approprié pour les régressions de survie mais SPSS réussit généralement à faire la même chose. Stata a des modèles de risques simultanés (competing risks): équivalent à logistique multinomiale de survie.
- Pour les analyses multi-niveaux, les logiciels spécifiques sont HLM, MLWin, M+ et R et dans une moindre mesure Lisrel, Stata, SPSS, SAS.
- Le transfert de bases de données d'un logiciel à un autre est habituellement facile. Les logiciels spécifiques lisent les fichiers de SPSS, STATA, R ou SAS.

Conclusion

- Au départ, il peut être plus difficile de travailler sur les fichiers pour pouvoir faire les analyses appropriées, MAIS
- Le jeu en vaut la chandelle. Une fois la base de données créée, tout devient nettement plus simple.
- Ne pas oublier que la première étape est de décrire, de visualiser.