

Modélisation hiérarchique, pooling partiel et l'interrogation des requêtes virtuelles

Andrew Gelman

Département de Statistique et Département des Science Politique,
Columbia University, New York

En visite à ENSAE, Paris

Merci à David Park, Boris Shor, Joe Bafumi, Yair Ghitza, Yu-Sung Su, Ken Shirley, Sharad Goel, David Rothschild, Doug Rivers, Wei Wang, François Briatte

Paris Machine Learning Meetup, 16 oct 2013

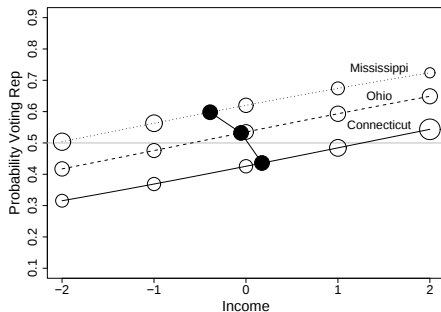


- ▶ On s'intéresse aux liens entre la politique et les inégalités économiques
- ▶ Nous avons étudié les votes des riches et pauvres dans les 50 États américains
- ▶ Je vous montrerai une série de modèles

États riches et pauvres, électeurs riches et pauvres

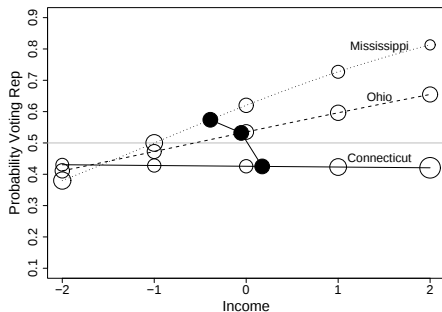
Le premier modèle

Varying-intercept model, 2000



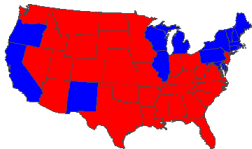
Le deuxième modèle

Varying-intercept, varying-slope model, 2000

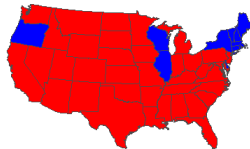


L'inférence ...

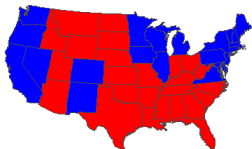
State winners in 2008 (rich voters only)



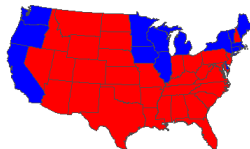
State winners in 2008 (rich Whites only)



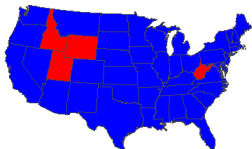
State winners in 2008 (middle-income voters)



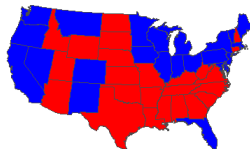
State winners in 2008 (middle-income Whites)



State winners in 2008 (poor voters only)



State winners in 2008 (poor Whites only)



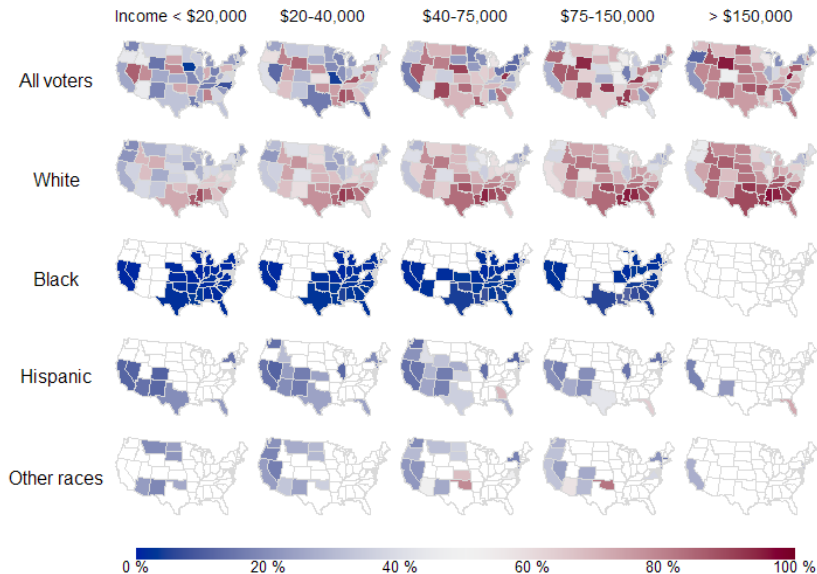
- ▶ Les critiques du blogueur “Daily Kos”:
 - ▶ Les critiques des inférences:

“While Gelman claims only the under-\$20K white demo went for Obama, the results were far different. Per the exit poll — real voters — Obama won all whites: 54-45 percent for those making under \$50K, and 51-47% for those making over \$50K. ... New Hampshire is solidly Blue unlike Gelman’s maps, 58-40 — one of the most obvious misses in Gelman’s analysis. ...”
 - ▶ Les critiques de la méthode:

“Gelman inexplicably avoids using exit poll data ... while exit polls have their own margin of errors and sample composition problems, they sure as heck beat anything done over the telephone.”

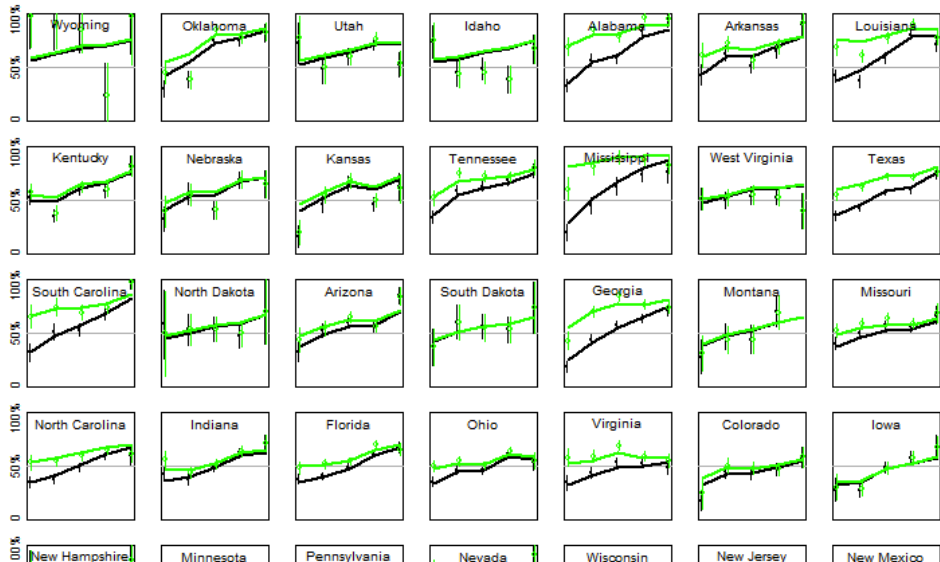
Après le changement du modèle

Did you vote for McCain in 2008?

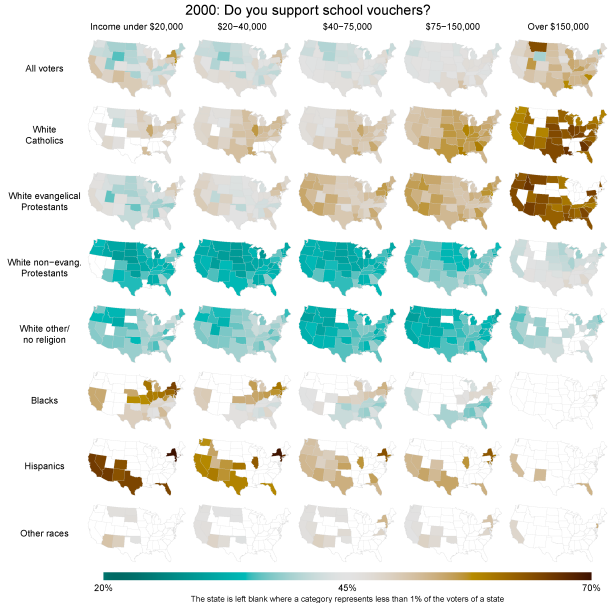
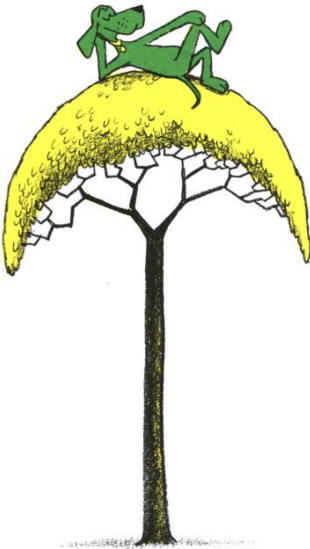


Un graphe pour comprendre et critiquer nos inférences

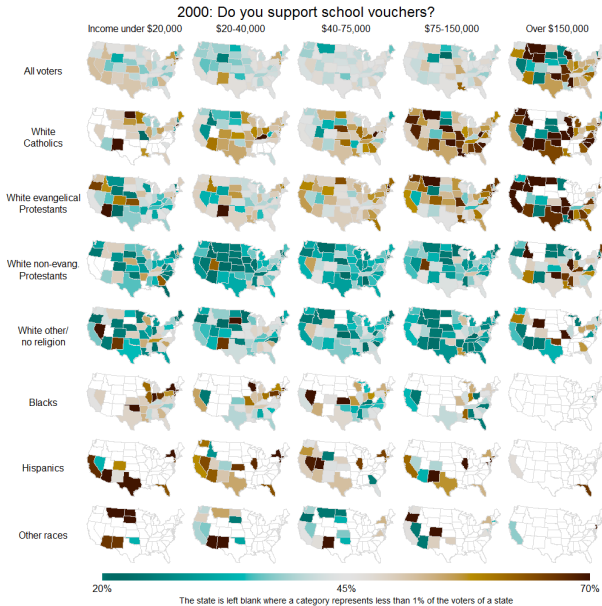
2008 election: McCain share of the two-party vote in each income category within each state among all voters (black) and non-Hispanic whites (green)



Les attitudes sur le paiement pour les “chèques-éducation”

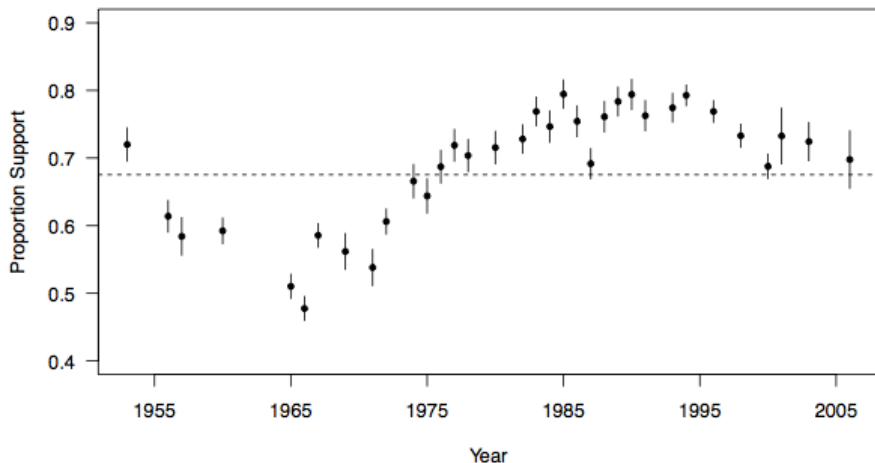


Les données ne parlent pas toujours d'elles-mêmes

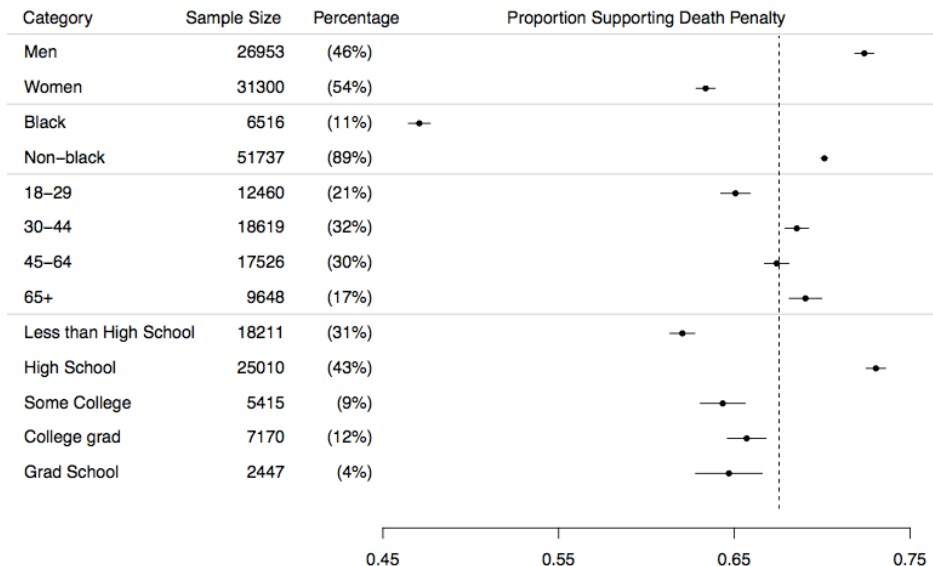


L'évolution des attitudes sur la peine de mort

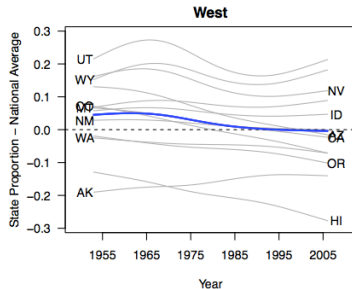
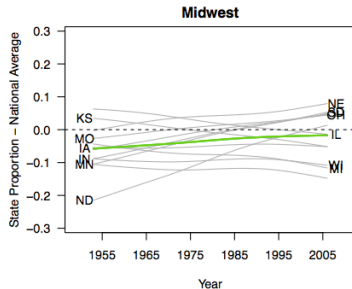
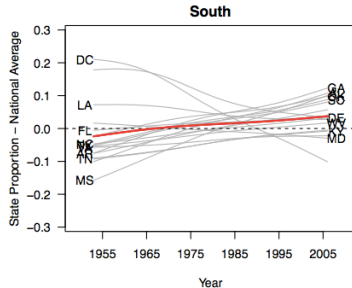
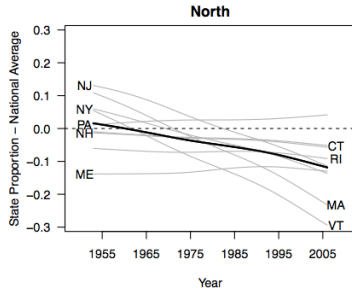
Proportion of Support by Year



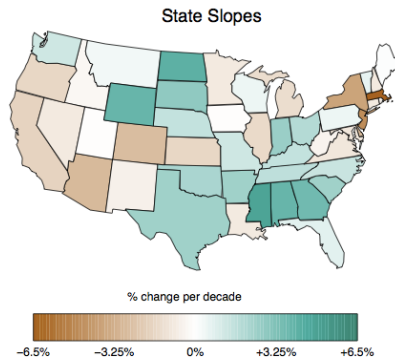
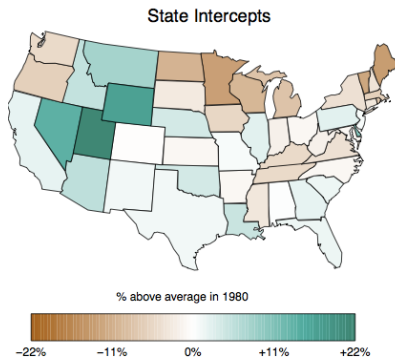
Par démographie



Tendances dans les États

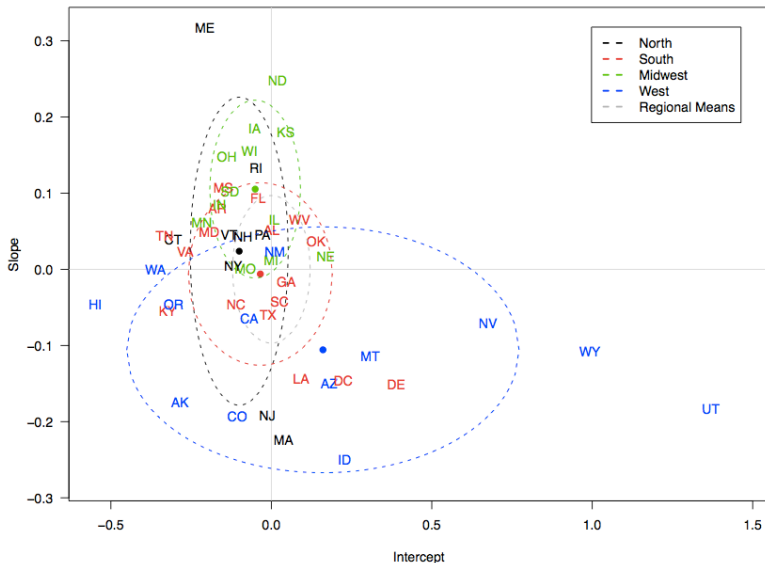


Les moyennes et les tendances

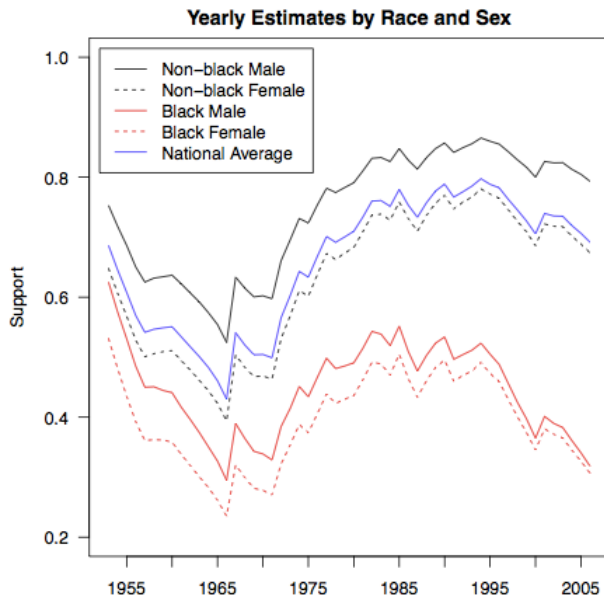


Variation non incluse dans le modèle

Slopes vs. Intercepts by State (within region): State-specific random effects

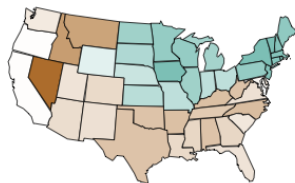


Tendances selon l'ethnicité et le sexe

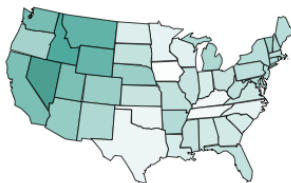


Comparaison par le niveau d'éducation

Degree: Less than High School



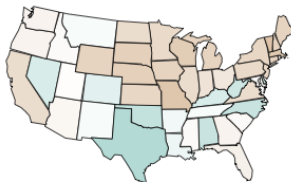
Degree: High School



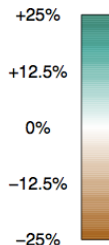
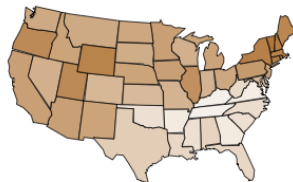
Degree: Some College



Degree: College Grad

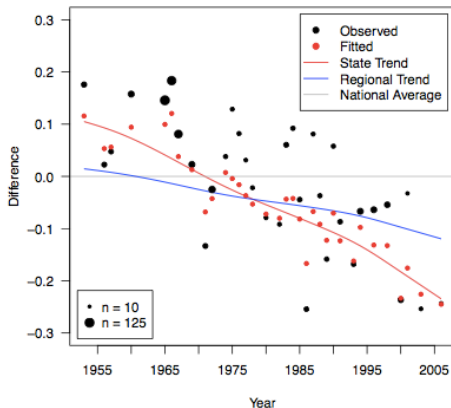


Degree: Grad School

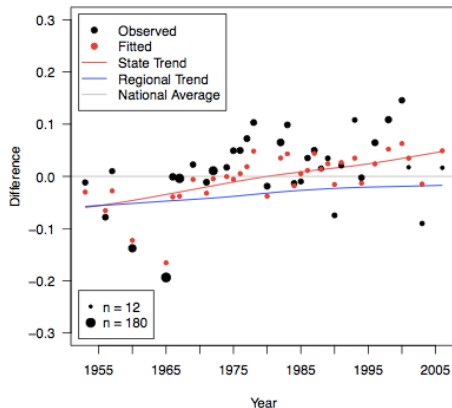


Focus sur deux États (après avoir soustrait la moyenne nationale): Les données et les inférences

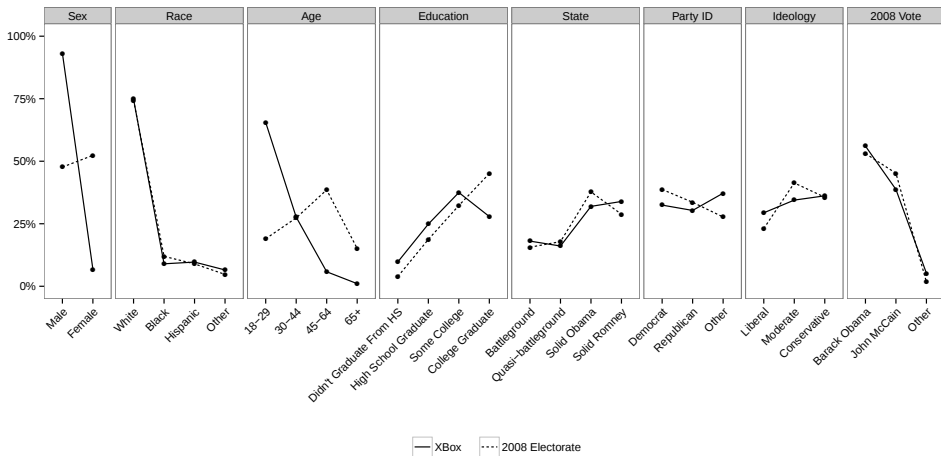
MA (North); Mean yearly sample size = 43



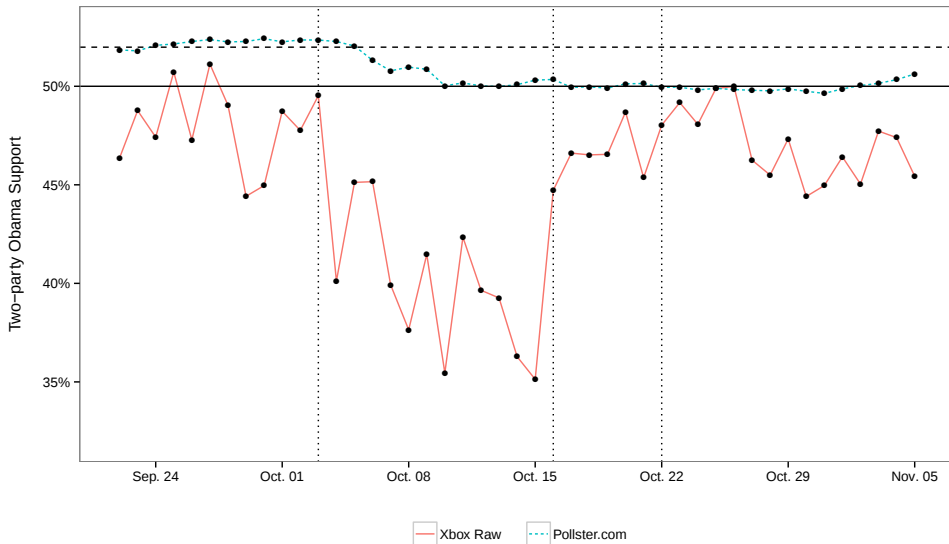
OH (Midwest); Mean yearly sample size = 85



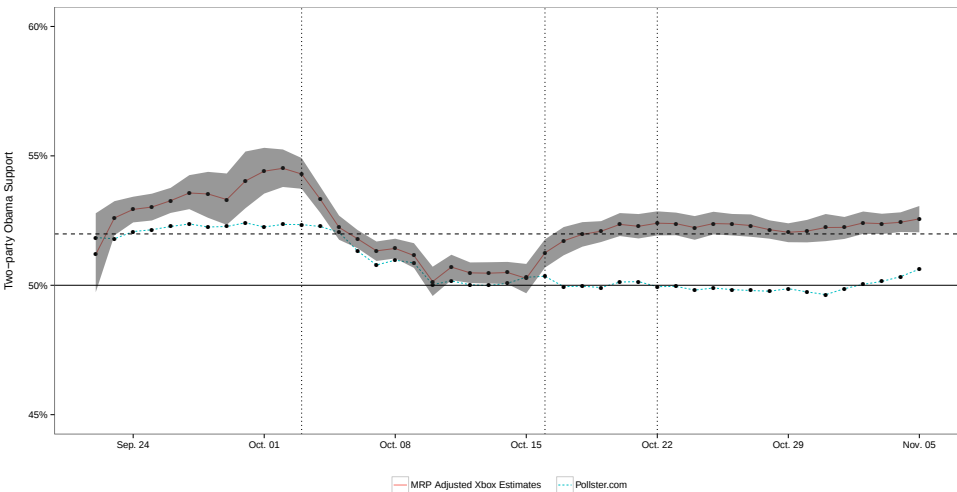
2012: Le sondage de Xbox



2012: Les données brutes

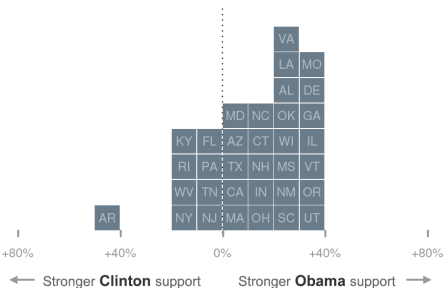


2012: Les données ajustées

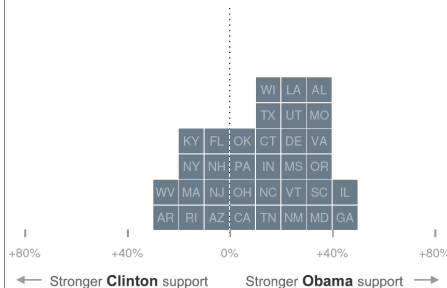


L'interrogation des requêtes virtuelles

Those with advanced education

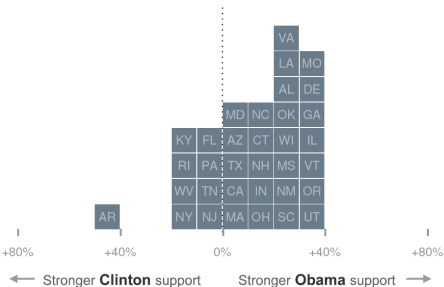


College graduates

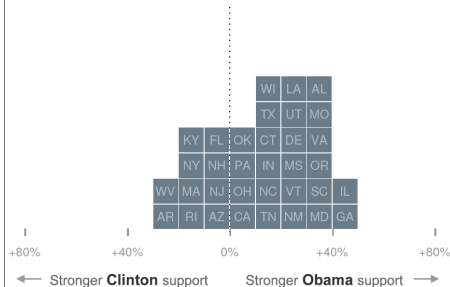


Le but du jeu

Those with advanced education



College graduates



C'est joli . . . mais nous voulons plus:

- ▶ Des divisions de populations plus petites
- ▶ Comparaisons à partir des données brutes
- ▶ Comprendre les effets du modèle et du pooling partiel



- ▶ Le modélisation hiérarchique, c'est une technique très flexible
- ▶ MRP
- ▶ L'interrogation des requêtes virtuelles
- ▶ On doit faire plus de recherche sur le modélisation bayésienne avec plusieurs facteurs (géographie, age, sexe, education, état civil, . . . , et toutes les interactions!)
- ▶ Aussi on a besoin de recherche sur les méthodes pour comprendre le modèle en contexte des données