



Synthèse sur la mesure - vers les analyses de prédiction

Cours Sol6210 Analyse quantitative avancée

© Claire Durand, 2023
Professeur titulaire,
Département de sociologie,
Université de Montréal




Ce qu'on a vu...

- Deux types d'analyse :
 - ▶ Relatives à la similarité entre les *réponses* (des individus, des cas, etc.): Analyses factorielles, des correspondances, en composantes principales, de classification sur les variables.
 - ▶ Relatives à la similarité entre les *cas*, individus, etc. dans les réponses données: analyses de classification, hiérarchique, en nuées dynamiques, mixte, etc.



Ce qu'on a vu...

- **Différences selon le type de variables:**
 - ▶ Variables de type quantitatif ou “pseudo-quantitatif”, soit binaires ou ordinales:
 - Analyse factorielle, en composantes principales
 - Analyses de classification des variables
 - Analyse de classification des cas pour variables continues
 - ▶ Variables de type nominal ou ordinal (peu de catégories):
 - Analyses de correspondances
 - Analyses de classification sur facteurs.
 - Analyses de classification two steps



Quand utilise-t-on lesquelles?

Rappel...



Les questions non abordées...

L'équivalence des mesures dans le temps et l'espace

- Si je veux créer des échelles ou des typologies, comment faire pour pouvoir répliquer sur d'autres échantillons différents dans le temps et l'espace?
- Est-ce que des différences dans le temps pour une mesure d'attitude sont des indications d'un manque de cohérence/fiabilité de la mesure ou des indications que des changements se sont produits. Comment différencier les deux?



Des questions?

- Si je veux utiliser des données provenant de différentes sources, mesurant le même construit mais de manière différente, est-il possible de mettre ces mesures sur la même échelle?
- Même mesuré de la même manière, est-il possible que certaines questions ne veuillent pas dire la même chose pour tout le monde?
- Les facteurs qui influencent les différences entre personnes peuvent-ils influencer l'équivalence des mesures?



La théorie de réponse aux items

- Permet de mettre des mesures différentes sur une même métrique.
- Permet de tenir compte du fait que la distance entre les points d'une échelle ordinaire n'est pas nécessairement la même pour tous les points.
- Permet de tenir compte du fait que chaque item n'a pas nécessairement la même importance dans la mesure du construit en soi et dans différents groupes.
- Permet d'estimer les différences dans les diverses versions d'un questionnaire en différentes langues.



D'autres questions

- Lorsque l'on utilise des méthodes exploratoires, on est limité
 - ▶ Par la procédure dont les paramètres de décision sont purement statistiques.
 - ▶ Par le choix d'un point de départ au hasard (analyse de nuées dynamiques).
 - ▶ Par l'impossibilité de modéliser une hiérarchie de variables latentes (facteurs et sous-facteurs, types et sous-types)



Les analyses de variables latentes

On renverse le “fardeau de la preuve”

- **Équations structurelles: Lisrel, Mplus, Amos, Stata, EQS, R,...**
 - ▶ Le modèle de mesure (pour l’analyse factorielle):
 - ▶ On peut modéliser qu’un facteur général influence trois facteurs spécifiques, qui chacun influence un certain nombre de variables mesurées.
 - ▶ On peut modéliser les variables complexes qui sont influencées par plus d’un facteur.
- **Analyses de classes latentes: Latent Gold,...**
 - ▶ Similaire: possibilité de déterminer d’avance de quoi aura l’air la typologie.



Les réseaux neuronaux

Algorithme de Kohonen

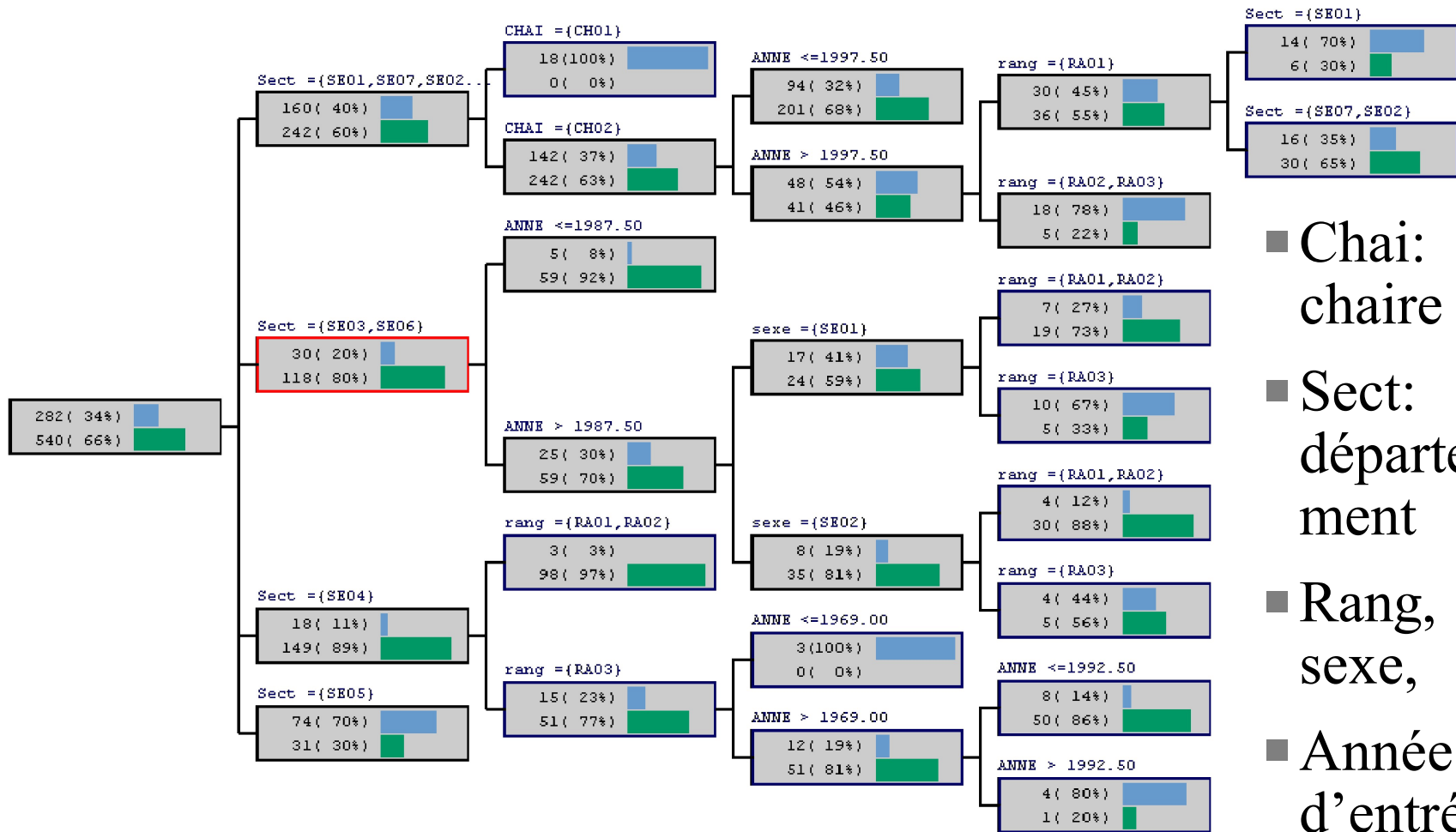
- Version plus sophistiquée des analyses de classification
 - ▶ La procédure apprend en classifiant. C'est un “précurseur” de l'intelligence artificielle (on parle de réseaux neuronaux).
- On se retrouve avec un certain nombre de “patrons” de réponse, de manières de répondre, et on peut ensuite les regrouper selon des similitudes.



Les analyses de segmentation

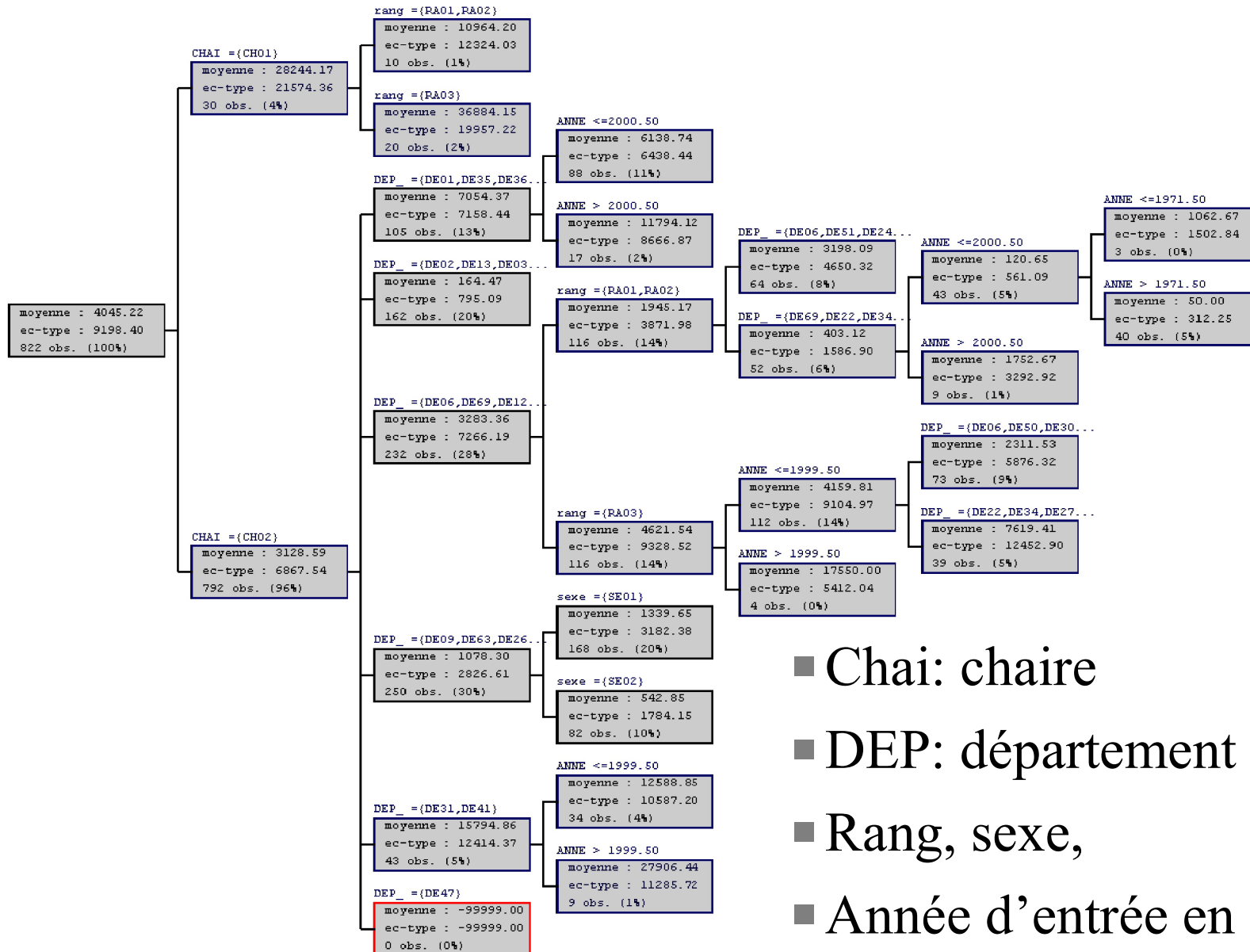
- On divise l'ensemble des personnes/cas selon leurs caractéristiques/réponses à un certain nombre de variables jusqu'à l'obtention de sous-groupes homogènes qui ne peuvent plus être subdivisés.
 - ▶ Chaid : SPSS
 - ▶ Analyses de segmentation pour variables catégorielles et pour variables quantitatives: SPAD

Présence prime salariale Université



- Chai: chaire
- Sect: département
- Rang, sexe,
- Année d'entrée en fn.,

\$ Prime salariale Université



- Chai: chaire
- DEP: département
- Rang, sexe,
- Année d'entrée en fn.,

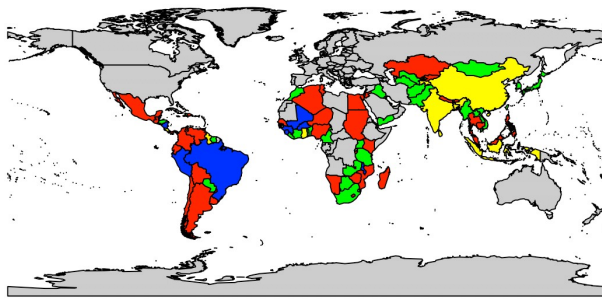


Les analyses de classification longitudinales

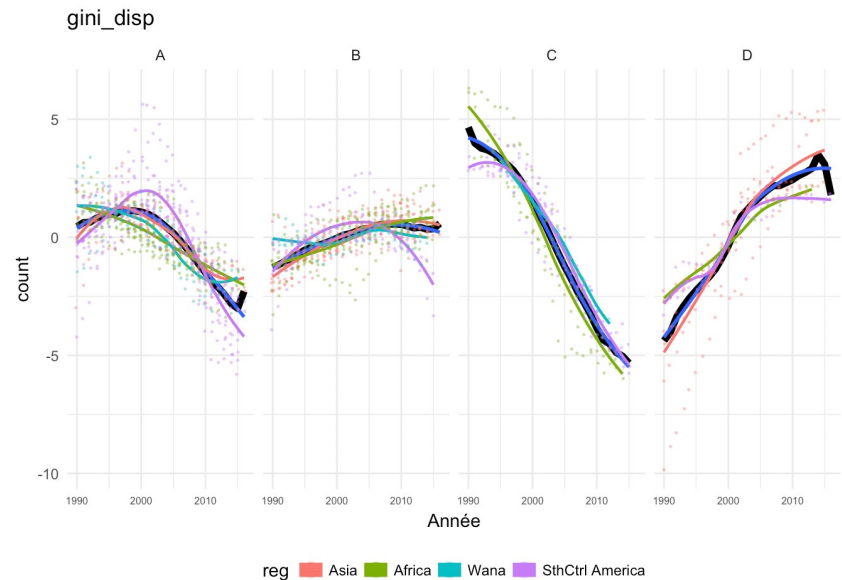
- Données pour 98 pays de 1995 à 2016.
- Analyses de type nuées dynamiques longitudinales.
- La question: Peut-on regrouper les unités (des pays dans notre cas) en fonction de l'évolution de la valeur de certaines variables?
 - ▶ Est-ce que c'est la tendance – augmentation, diminution, stabilité – qui a un impact plutôt que seulement le niveau à chaque année par exemple?

Gini_disp index - shape

Estimate of inequality in equivalized household disposable (post-tax, post-transfer) income; the clustering maximises the shape of change over time, independently from the level.



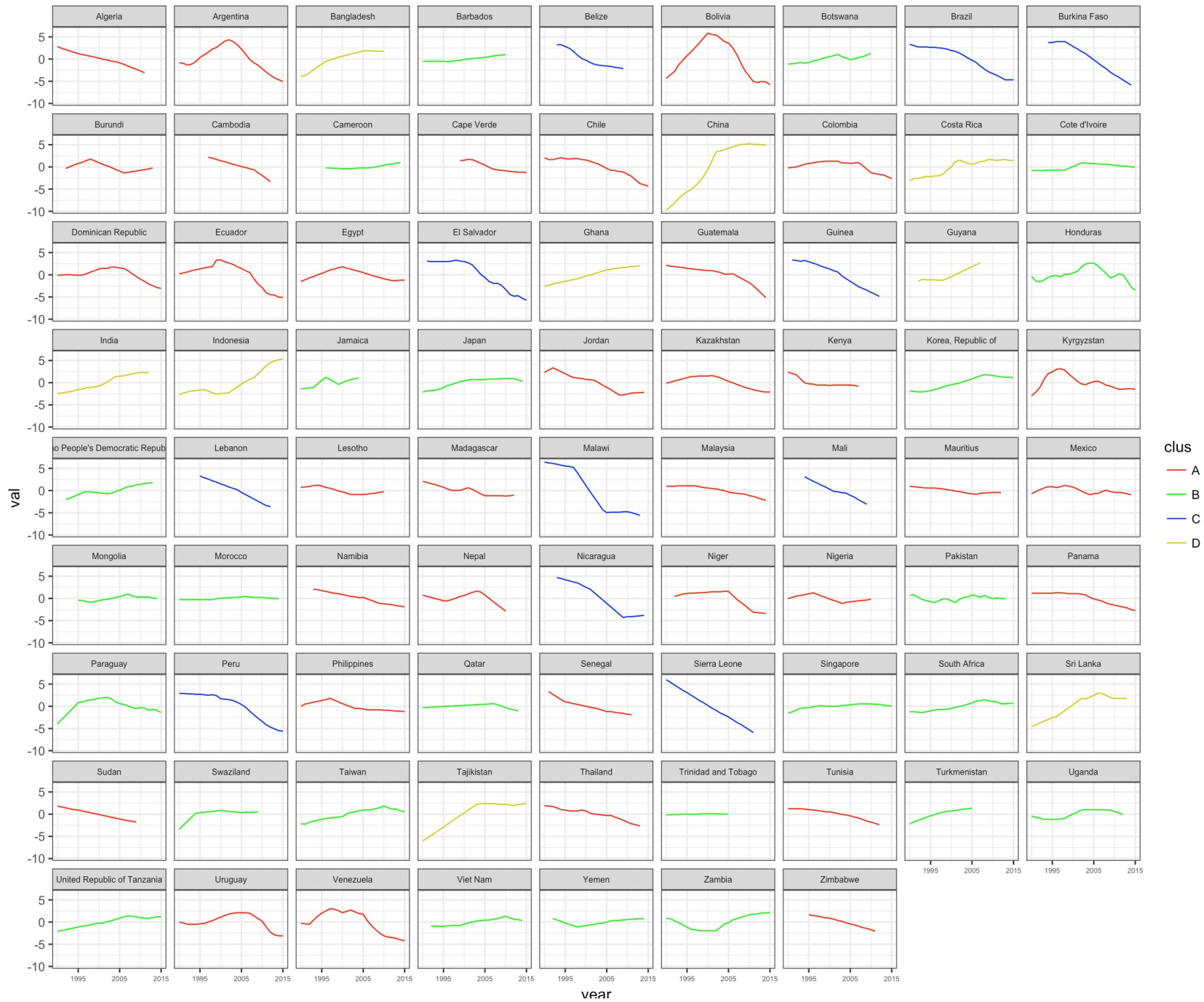
A-Rouge, B-Vert, C-Bleu, D-Jaune



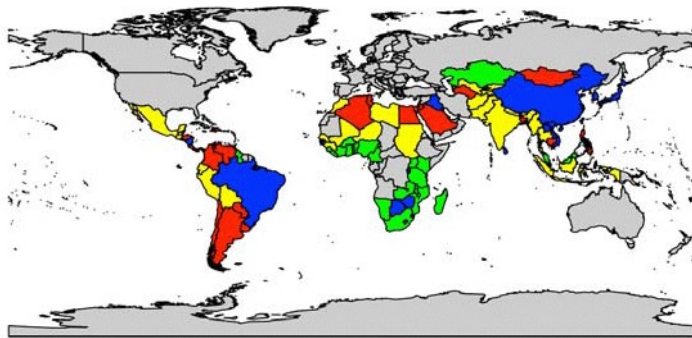
Latin America: Quadratic trend or decreasing (blue & red).

Africa & Wana: mixed, but not increasing.

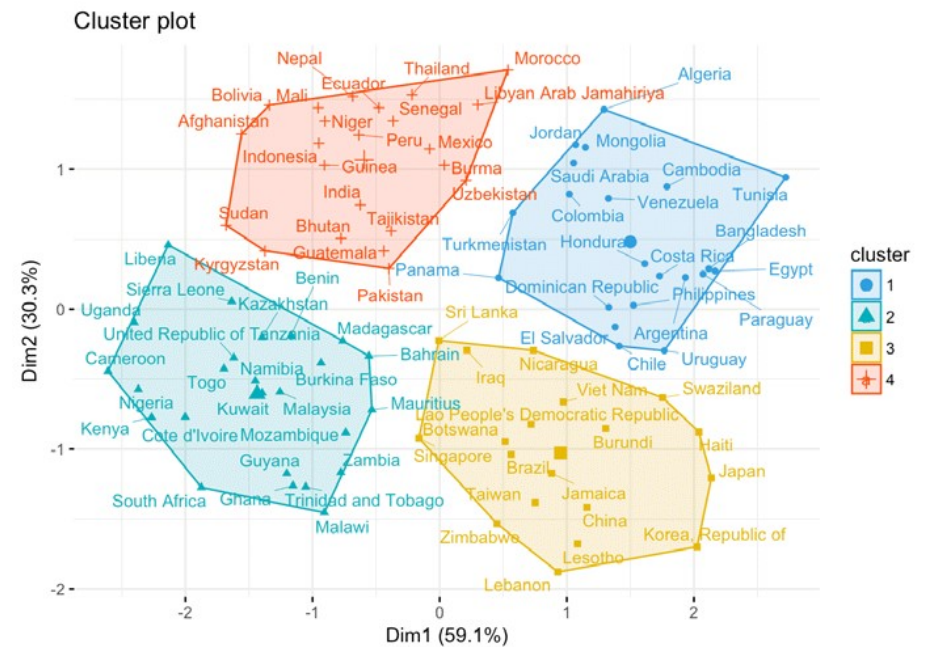
Asia: mostly increasing (Yellow)



Diversity - ethnic & religious

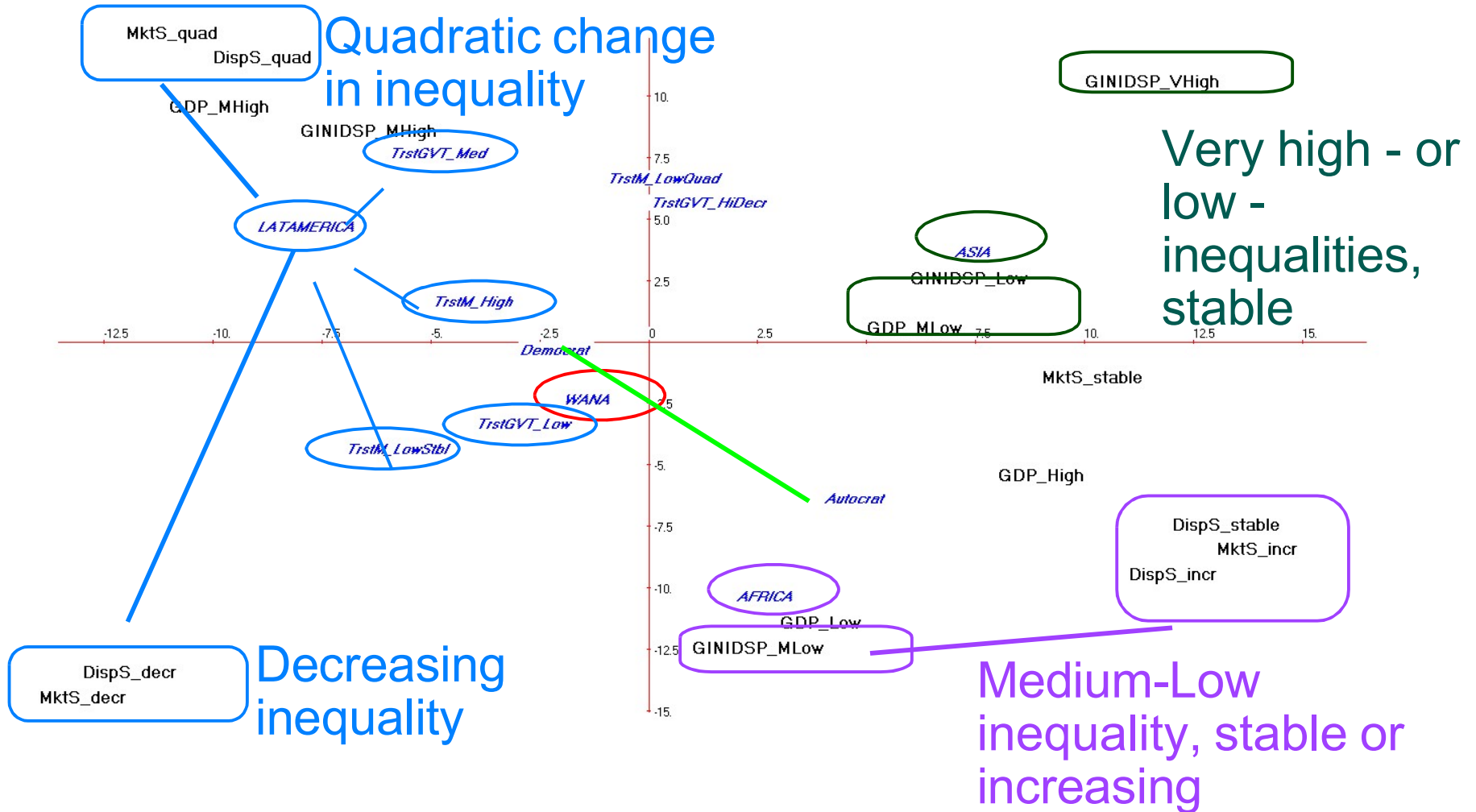


1-Rouge, 2-Vert, 3-Bleu, 4-Jaune



Variation within regions:
 Red: Low diversity; Green: High diversity
 Blue: Low ethnic, high religious
 Yellow: High ethnic, low religious

Relationship between clusters of economic situation & clusters of trust?



The graph shows no relationship bw trust & economic situation



L'analyse discriminante...

- Parfois regroupée avec les analyses de type factoriel mais...
- En quelque sorte, il s'agit de la relation entre une analyse factorielle et des types
 - Peut-on trouver une/des combinaisons des variables indépendantes qui permettront de bien prédire l'appartenance des individus à des groupes/types?



Pourquoi utiliser ces analyses?

- On peut découvrir des patrons de relations, des pistes auxquels on ne s'attendait pas.
- On assimile mieux l'ensemble complexe des relations entre les diverses variables.
- Elles nous guident dans le regroupement ou la sélection de variables qui seront retenues pour les analyses de type prédictif.



Pourquoi utiliser ces analyses?

- Parce que les analyses de type prédictif (régressions de divers types) requièrent la plus grande indépendance possible entre les variables indépendantes.
 - ▶ → Éviter la collinéarité.
- Parce qu'on aura créé de nouvelles variables plus fiables ou dont, à tout le moins, on peut évaluer la fiabilité.
 - ▶ → Permettre de mettre en perspective les conclusions statistiques liées aux relations entre les variables.



Conclusion

Avant tout...

- Il faut régler le problème du modèle de mesure.
- Les analyses présentées permettent de procéder à cette première étape.
-
- Et il faudra reparler de la greffe de fichiers...