Introduction à la démarche de recherche et à l'ensemble des méthodes

Claire Durand, professeur
Cours Sol 6210, Analyse quantitative avancée

BUT DE LA RECHERCHE

- THÉORIE EXISTANTE: vérifier, valider
- THÉORIE À ÉLABORER: proposer, élaborer d'abord, début de validation, vérification ensuite
- DE LA THÉORIE DÉCOULENT...
  - Les concepts
  - Les définitions
  - L'opérationnalisation
  - Les mesures

De la théorie découlent

Des concepts à mesurer

- Il faut donc élaborer, puis valider, les mesures des concepts:
- Nos outils: les statistiques descriptives
  - UNIVARIÉES (fréquences, statistiques de distribution)
  - MULTIVARIÉES :
    - *Analyse en composantes principales/ Analyse factorielle -- exploratoire ou confirmatoire*
    - *Analyse des correspondances*
    - *Analyse de classification*
    - Autres:
      - *Théorie de réponse aux items*
      - *Multidimensional scaling*

De la théorie découlent

- Des propositions concernant les relations entre les concepts.
- Des hypothèses concernant les relations entre les variables.
- Nos outils:
  - **LES STATISTIQUES BI-VARIÉES (Chi Carré, Test F,...)**
  - **LES STATISTIQUES EXPLICATIVES, "PRÉDICTIVES":**
    - AUCUNE VARIABLE DÉPENDANTE:
      - Modèles loglinéaires non-hiérarchiques
    - UNE VARIABLE DÉPENDANTE:
      - Régressions linéaires multiples
      - Régressions logistiques dichotomiques, polytomiques
      - Analyse discriminante
      - Régressions de survie, séries chronologiques,…

- UNE VARIABLE DÉPENDANTE (suite):
  - Analyse multi-niveaux, (hiérarchique), qui peut être longitudinale
  - Modèles loglinéaires hiérarchiques
  - Modèles logit et probit
- PLUSIEURS VARIABLES DÉPENDANTES:
  - Analyse canonique
  - MANOVA, MANCOVA

- **De la théorie découlent des modèles de relations entre l'ensemble des variables.**
  - Des techniques statistiques spécifiques à la validation de modèles.
    - Équations structurales (LISREL, EQS), systèmes d'équations simultanés.
    - Analyses de classes latentes

PROCESSUS

- À partir de la théorie, cueillette, construction des données (provenant de diverses sources – questionnaires, discours, données institutionnelles, …)
- Transformation de l'information qualitative en données numériques (chiffres!) qui constituent une approximation de la réalité.
- **Processus:**
  - PHASE 1A) «sentir les données», explorer statistiques descriptives uni et multivariées.
  - PHASE 1B) validation des mesures
  - PHASE 2) analyses bi-variées: analyse des relations et de l'enchevêtrement des relations.
  - PHASE 3) la prédiction d'une ou de plusieurs variables dépendantes
  - PHASE 4) l'élaboration, la vérification de modèles de relation, l'« explication», la compréhension.

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES

ANALYSE EN COMPOSANTES PRINCIPALES ET ANALYSE FACTORIELLE

UTILITÉ:

- données métriques, ordinales ou dichotomiques
- permet de connaître le "patron", la structure sous-jacente aux relations entre les variables
- sert à éliminer les mesures qui n'appartiennent pas à un seul concept et à orienter la création de mesures fidèles des concepts.

VARIANTES

- Analyse en composantes principales:
  - prend en compte l'ensemble de la variance des variables dans l'analyse.
  - donne une solution unique et des facteurs orthogonaux (non-corrélés)
- Analyse factorielle
  - prend en compte uniquement la variance commune à l'ensemble des variables
  - donne la "meilleure décomposition statistique" en facteurs (plusieurs solutions possibles).
  - plusieurs modes d'extraction permettent de tenir compte de la distribution des données et de la volonté de maximiser certains aspects plutôt que d'autres
  - les facteurs peuvent être corrélés (tient compte de la relation entre les mesures)

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES (suite)

ANALYSE EN COMPOSANTES PRINCIPALES ET ANALYSE FACTORIELLE (suite)

INFORMATION:

- Plusieurs indices permettent de déterminer
  - le nombre de facteurs le plus approprié (valeur propre - critère de Kaiser, coude de Cattell - test de l'éboulis)
  - l'adéquation de la solution factorielle (variance expliquée, KMO)
  - l'appartenance de chaque variable à la solution (qualité de la réprésentation, saturations factorielles)

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES (suite)

ANALYSE FACTORIELLE DES CORRESPONDANCES

UTILITÉ:

- données nominales ou ordinales
- permet de visualiser les catégories selon la similarité des réponses.
- sert à déterminer s'il y a lieu de créer des regroupements de catégories et sert à orienter la sélection des variables.

VARIANTES:

- Simple (2 variables, multiples catégories)
- Multiple (plusieurs variables)

INFORMATION:

- **plusieurs indices permettent de déterminer *:***
  - le nombre de facteurs nécessaires pour expliquer la majeure partie de l'information, le cas échéant (valeurs propres)
  - l'adéquation d'une solution factorielle: Peut-on regrouper les catégories en préservant suffisamment l'information pertinente?
  - l'appartenance de chaque catégorie de variable à un facteur.
  - à quelle catégorie ou regroupement de catégories (profil) ressemblent ceux qui n'ont pas répondu.

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES (suite)

ANALYSE DE CLASSIFICATION HIERARCHIQUE (CLUSTER) ET ANALYSE DE NUÉES DYNAMIQUES (QUICK CLUSTER)

UTILITÉ :

- traditionnellement traite des données métriques.
- analyse “à la française” sur les scores factoriels obtenus suite à une analyse de correspondance.
- l'analyse cherche à regrouper les individus (ou les variables) en fonction de la similitude entre eux, ce qui donne des profils-type.
- l'analyse regroupe les individus (ou les variables) en classes selon divers critères, diverses mesures de distance entre les cas et diverses façons de les regrouper.

VARIANTES :

- Analyse portant sur les individus
  - permet de regrouper les "cas" en groupes ayant des caractéristiques similaires.
  - ces regroupements facilitent la constitution et la validation de typologies.
- Analyse portant sur les variables
  - constitue l'ancêtre de l'analyse factorielle et donne des résultats similaires à celle-ci.
  - peut être utilisée de préférence à l'analyse factorielle quand les prérequis à l'utilisation de celle-ci ne sont pas remplis (multi-normalité, 10 cas par variable...).

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES (suite)

ANALYSE DE CLASSIFICATION HIERARCHIQUE ET ANALYSE DE NUÉES DYNAMIQUES (CLUSTER) (suite)

VARIANTES (suite):

- Hiérarchique: regroupement systématique en fonction de mesures de distance entre les points et de critères de regroupement
- Nuées dynamiques: processus itératif de classification
- Mixte: mélange des deux procédures précédentes.

INFORMATION :

- nombre optimal de classes
- distance entre les classes
- appartenance de chaque cas à une classe

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES (suite)

Analyses de type confirmatoire, nouveaux modèles d'analyse:

- Pour les modèles de mesure de type "Analyse factorielle": Équations structurelles (modèle de mesure)
- Pour les modèles de types 'classification hiérarchique": Analyse de classes latentes

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES

THÉORIE DE RÉPONSE AUX ITEMS

UTILITÉ :

- complémentaire à l'analyse factorielle, elle permet de créer des échelles standardisées
- les scores produits ne sont pas les scores additifs traditionnels; ce sont des scores différents selon les patrons de réponse; ils tiennent compte des profils de réponse différents selon les groupes.

VARIANTES :

- modèles à un, deux ou trois paramètres (paramètres de discrimination, de difficulté et de hasard)
- modèles pour bâtir des tests d'habileté (avec une réponse valide)
- modèles pour les échelles de mesure des attitudes avec une ou plusieurs catégories (Graded model de Samejima)

PHASE 1: EXPLORATION DES DONNÉES ET VALIDATION DES MESURES (Nouveaux modèles)

THÉORIE DE RÉPONSE AUX ITEMS (suite)

INFORMATION :

- Jusqu'à quel point une question, souvent appelée un item dans cet univers, permet de discriminer un individu d'un autre sur une échelle mesurant le positionnement des individus sur un concept, ceci indépendamment des questions choisies pour mesurer le concept.
- Jusqu'à quel point les individus ayant un patron de réponse spécifique sont plus ou moins "habiles" dans une compétence ou "favorables" dans leur positionnement par rapport à une attitude déterminée.

Synthèse

En première phase, on cherche souvent à synthétiser un ensemble plus ou moins disparate d'informations en un nombre moins élevé et gérable de mesures.

- En créant des variables composites.
- En créant des classes de répondants.
- En combinant ces deux familles de méthodes.

PHASE 2: LES RELATIONS BI-VARIÉES, L'ENCHEVÊTREMENT DES RELATIONS

UTILITÉ :

- Les relations bi-variées (analyses de variance, tableaux de contingence, régressions simples) permettent de faire un portrait détaillé de l'ensemble des relations entre les variables et d'en tirer les premiers éléments de sens. Ceci permet de faire une première synthèse des relations.

INFORMATION :

- Donne les informations sur les relations "nettes" entre deux variables; l'ensemble des relations bi-variées (parfois il est aussi nécessaire de vérifier certaines relations tri-variées) permet de voir s'il faut soupçonner l'existence de relations factices (spurious correlation), de variables modératrices, de relations "explicatives".

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES

RÉGRESSIONS LINEAIRES MULTIPLES

UTILITÉ :

- Données de forme métrique ou "pseudo-métrique (0,1)"
- Sert à prédire la valeur d'une variable pour un individu donné à partir d'une ou plusieurs variables ou groupes de variables.

VARIATIONS relatives au **mode d'entrée** des variables indépendantes:

(vaut pour tous les types de régression)

- Régression standard:
  - Toutes les variables prédictrices d'intérêt sont entrées dans l'analyse en même temps
  - Permet de connaître la contribution unique de chaque variable, compte tenu de la contribution partagée avec les autres variables.
- Régression hiérarchique/séquentielle:
  - L'ordre d'entrée des variables est déterminé théoriquement.
  - Permet d'estimer la contribution d'une variable ou d'un groupe de variables au-delà de la contribution des variables déjà dans l'équation.

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES (suite)

RÉGRESSIONS LINEAIRES MULTIPLES (suite)

VARIATION dans les modes d'entrée: (suite)

Régression statistique:

- L'ordre d'entrée des variables est déterminé statistiquement
- Permet de trouver un bon sous-ensemble de variables prédisant la variable dépendante, ceci de façon automatisée en fonction de critères purement statistiques.

INFORMATION:

- Plusieurs indices permettent de déterminer
  - la contribution de chaque variable à la prédiction (B, Beta, test T corrélation semi-partielle)
  - la proportion de variance expliquée par une variable ou un groupe de variables (r^2, Δr^2)
  - la justesse de la prédiction selon les valeurs de la variable dépendante (analyse des résidus)

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES (suite)

RÉGRESSIONS LOGISTIQUES

UTILITÉ :

- variable dépendante de type nominal, multinominal ou ordinal et variables indépendantes de diverses formes
- permet de prédire la probabilité d'appartenir à une catégorie d'intérêt de la variable dépendante (plutôt qu'à la catégorie de référence)

VARIANTES :

- Tout comme la régression linéaire, le mode d'entrée des variables peut être standard, hiérarchique ou statistique.
- Régression logistique dichotomique
  - régression où la variable dépendante est de forme (0,1), et donc, dénote l'absence ou la présence d'une qualité, d'un événement.
  - permet de prédire la probabilité d'appartenir à la catégorie 1 plutôt que de ne pas y appartenir, compte tenu des variables dans l'équation.

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES (suite)

RÉGRESSIONS LOGISTIQUES (suite)

- Régression logistique polytomique
  - régression où la variable dépendante est de forme multinominale, soit la présence de *plus d'une* qualité comparée à l'absence de toute qualité. La variable dépendante peut aussi être de forme ordinale.
  - permet de prédire la probabilité d'appartenir à l'une ou l'autre des catégories d'intérêt plutôt que d'appartenir à la catégorie de référence.

INFORMATION:

- Plusieurs indices permettent d'estimer:
  - la contribution de chacune des variables à la prédiction globale (χ^2 de Pearson).
  - la contribution de chacune des catégories des variables indépendantes à la prédiction de chacune des catégories d'intérêt (Test Wald).
  - le rapport de cote, c'est-à-dire la probabilité qu'un événement survienne étant donné les informations possédées sur les variables indépendantes.
  - la justesse de la prédiction (analyse des résidus).
  - l'adéquation du modèle (χ^2 de maximum de vraisemblance).

PHASE 3: LA PRÉDICTION D'UNE OU PLUS VARIABLES DÉPENDANTES - liées au temps

RÉGRESSIONS DE SURVIE, SÉRIES CHRONOLOGIQUES, ANALYSES MULTI-NIVEAUX LONGITUDINALES, ANALYSES DE TRAJECTOIRES

UTILITÉ :

- Variable dépendante se modifiant avec le temps:
  - Les observations ne sont pas indépendantes entre elles et il y a plusieurs observations pour un même individu.
  - Il peut aussi y avoir une série d'informations similaires pour un certain nombre d'unités de temps
- Permet d'estimer la probabilité cumulative qu'un événement survienne et d'examiner quelles sont les variables qui prédisent mieux cette probabilité.
- Permet d'estimer le meilleur modèle de l'évolution d'une variable dans le temps.
- Permet d'estimer des groupes de trajectoires homogènes.
- Permet d'utiliser l'ensemble de l'information disponible même si certaines observations n'ont pas pu être faites pendant toute la période d'observation.

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES - liées au temps

RÉGRESSIONS DE SURVIE, SÉRIES CHRONOLOGIQUES, ANALYSES MULTI-NIVEAUX LONGITUDINALES, ANALYSES DE TRAJECTOIRES (suite)

VARIANTES :

- **Régression de survie (on parle aussi d'analyse de transition)**
  - Analyse des "tables de survie", i.e. le temps nécessaire pour qu'un événement survienne (obtention d'un emploi, changement d'emploi, naissance d'un enfant, etc.)
  - Les prédicteurs peuvent varier ou non avec le temps
- **Séries chronologiques**
  - Analyse une série de mesures prises à un intervalle de temps fixe, comme, par exemple, des sondages électoraux faits à chaque jour d'une campagne électorale, des décisions budgétaires faites à chaque année pendant un certain nombre d'années, l'évolution des taux de chomage, d'homicides. A noter que l'effet d'événements sur des séries chronologiques peut être immédiat ou différé sur une certaine période.
- **Analyse longitudinale multi-niveaux**
  - Modèles développés récemment qui ont l'avantage d'être plus souples. Permettent d'estimer la valeur d'une variable variant dans le temps et son évolution en tenant compte de variables liées tant au niveau 1 (variable dépendante variant dans le temps) qu'au niveau 2 (par exemple individu ayant des caractéristiques spécifiques).

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES - liées au temps

RÉGRESSIONS DE SURVIE, SÉRIES CHRONOLOGIQUES, ANALYSES MULTI-NIVEAUX LONGITUDINDALES, ANALYSES DE TRAJECTOIRES (suite)

- **Analyse de trajectoires ou de classification longitudinale**
  - Modèles très récents permettant d'estimer des typologies de trajectoires, c'est-à-dire des regroupements de personnes similaires quant à l'évolution d'une variable dans le temps (voir parcours de délinquance, par exemple).

INFORMATION:

- Quels sont les facteurs qui expliquent l'évolution de la V.D. d'un état à un autre à chaque intervalle de temps étant donné son état à l'intervalle de temps précédent? (Régression de survie)
- Quels sont les événements qui expliquent l'évolution d'une série chronologique (ex. Le débat télévisé en campagne électorale a-t-il influencé l'évolution subséquente de l'intention de vote?)
- L'évolution de la variable dans le temps est-elle la même pour tous les individus?
- Peut-on regrouper les individus selon un nombre fini de profils en fonction de l'évolution de la VD dans le temps? (Carrières de délinquance, carrières comparées des hommes et des femmes dans une entreprise, processus d'intégration des immigrants.)

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES

ANALYSE MULTI-NIVEAUX (aussi appelée HIÉRARCHIQUE)

UTILITÉ :

- Variable dépendante de type métrique, qualitative, etc.
- Permet de distinguer le niveau où l'effet se produit, d'utiliser des variables indépendantes mesurées à différents niveaux (individu, équipe de travail/classe/quartier, organisation)
- particulièrement utile dans les études où divers niveaux sont en cause de façon claire (élèves, classes, écoles; employés, équipes de travail, usines, organisations)

INFORMATION :

- contribution de chaque niveau d'analyse à la variation de la variable dépendante
- Contribution de facteurs aux différents niveaux dans l'explication de la variance à chaque niveau.

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES: AUTRES PROCÉDURES D'ANALYSE (non vues dans le cadre du cours)

MODÈLES LOGLINÉAIRES, PROBIT, LOGIT:

UTILITÉ :

- Données uniquement sous forme nominale, multinominale ou ordinale
- permet d'estimer l'adéquation d'un modèle sans nécessairement présumer de l'existence d'une variable dépendante

VARIANTES :

- Modèles loglinéaires "ordinaires"
  - modèles où il n'y a pas de variable dépendante
  - permettent d'estimer les relations entre une certain nombre (pas trop élevé) de variables nominales, multinominales ou ordinales.
- Modèles loglinéaires hiérarchiques
  - modèles où il y a une variable dépendante
  - permettent d'estimer la contribution des variables indépendantes (nominales, multinominales ou ordinales) à la prédiction d'une variable dépendante de même forme.

Note: On classe habituellement sous le groupe des modèles loglinéaires les régressions logistiques et les modèles LOGIT et PROBIT puisque tous ces types d'analyse utilisent des transformations logarithmiques

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES: AUTRES TYPES D'ANALYSE

(non vues dans le cadre du cours)

ANALYSE DISCRIMINANTE :

UTILITÉ :

- Variable dépendante de forme nominale ou multinominale (comme la régression logistique) et variables indépendantes de forme métrique ou "pseudo-métrique".
- Permet de prédire l'appartenance à l'une ou l'autre des catégories de la variable dépendante par une fonction discriminante qui résulte de la meilleure combinaison des variables indépendantes.
- Les prérequis pour l'utilisation de cette technique sont plus sévères que pour la régression logistique (multinormalité et restriction sur la forme des variables indépendantes).

INFORMATION :

- Adéquation de la classification de la variable dépendante par la fonction discriminante.
- Meilleure(s) combinaison(s) linéaire(s) des variables indépendantes susceptible(s) de discriminer entre les catégories de la variable dépendante.

PHASE 3: LA PRÉDICTION D'UNE OU DE PLUSIEURS VARIABLES DÉPENDANTES: AUTRES TYPES D'ANALYSE

(non vues dans le cadre du cours)

ANALYSES DE SEGMENTATION (CHAID, i.e., Chi square automatic interaction detection). Voir SPSS et SPAD.

ANALYSES AVEC PLUSIEURS VARIABLES DÉPENDANTES

- Analyse de variance multivariée (MANOVA OU MANCOVA)

ANALYSE CANONIQUE:

- Relation entre deux analyses factorielles.

PHASE 4: L'ÉLABORATION, LA VÉRIFICATION DE MODÈLES DE RELATION, L'"EXPLICATION"

ÉQUATIONS STRUCTURELLES ou simultanées (LISREL, EQS)

UTILITÉ :

- Permet de vérifier s'il est plausible que les relations constatées entre les variables dans l'échantillon correspondent au modèle théorique de relation posé.
- Les variables doivent généralement être de type métrique ou ordinale pour les modèles courants (traités avec LISREL, EQS, STATA ou le module AMOS de SPSS). Ce type de programme permet aussi de produire des analyses descriptives poussées et différents types de matrices (de covariance, de corrélations, de corrélations polychoriques, polysérielles, etc.).

INFORMATION :

- donne plusieurs indices d'adéquation globale du modèle.
- donne les estimés des paramètres donnant la force des relations entre les indicateurs et les variables latentes et entre les variables latentes elles-mêmes.
- donne des indices de l'adéquation de chacun des éléments du modèle.
- donne des indications sur la variance expliquée à chaque "niveau" de relation (indicateur vs variable latente et équations de régression).