Note méthodologique : De Nvivo à FsQCA : comment utiliser des codages de verbatims pour rechercher des causalités complexes ?

Sébastien Brion

Note méthodologique

De Nvivo à FsQCA : comment utiliser des codages de verbatims pour rechercher des causalités complexes ?

Sébastien Brion, IREGE - 2014

I. Objectif méthodologique de la démarche

Ce document a pour but de décrire la démarche méthodologique à adopter pour transformer des données qualitatives textuelles (dans Nvivo par ex) en données exploitables pour tester des configurations en logique floue sur une variable expliquée (dans FsQCA[1]). L'objectif méthodologique du chercheur doit être de rechercher plusieurs configurations et combinaisons de variables pouvant expliquer un phénomène (une condition particulière).

Cette démarche ne consiste pas uniquement à quantifier des données qualitatives, mais tente également de déterminer/explorer les dimensions explicatives d'un phénomène, non pas en moyenne comme pourrait le faire une régression statistique classique mais en tentant au contraire de mettre en avant les configurations équiprobables (plusieurs combinaisons de variables peuvent expliquer un même phénomène) et les configurations « extrêmes » ou « *outliers* » qui sont généralement mises de côté par l'analyse statistique classique.

Les avantages de la démarche sont nombreux. Les données qualitatives textuelles ou graphiques sont très riches mais très difficiles à analyser lorsqu'elles sont disponibles en grande quantité. Cette démarche propose temporairement de les réduire à des données chiffrées pour les analyser et ensuite les réutiliser pour enrichir les résultats issus de cette réduction chiffrée.
La démarche vise ainsi à ne pas perdre le lien entre données chiffrées et leur contenu sémantique (textuelles, graphiques etc…). Réduire des textes en données chiffrées permet de les rendre comparable. Avant de recourir à cette méthode, le chercheur devra s'assurer qu'il existe une variable explicative unique dans son modèle (un phénomène unique à expliquer) permettant de comparer les configurations qu'il pourra mettre en évidence. En outre, cette démarche suppose que toutes les variables impliquées dans les configurations explicatives ont le même poids. Or, les démarches d'analyse qualitatives par codage thématique ont le plus souvent pour objectif de pondérer les discours des répondants et de mettre en évidence des thèmes récurrents et dominants. Si tel est l'objectif du chercheur, la méthode décrite ici ne lui sera d'aucune utilité dans la mesure où il sera justement question de transformer les concepts issus du codage thématique pour leur attribuer un même poids afin qu'ils puissent être analysés pour donner lieu à des configurations (seul le poids accordé par le répondant à un concept est pris compte). La démarche consistera à transformer les verbatims reliés à

[1] FsQCA est un logiciel de traitement de données basé sur l'algèbre booléen. Bien que très répandu, il n'est pas aujourd'hui le plus complet (cf. étude comparative de Thiem, A., & Dusa, A. (2013). "QCA: A package for qualitative comparative analysis", The R Journal 5(1): 1-11.). http://www.compasss.org/software.htm#fsQCA (lien de téléchargement consulté le 18/11/15)

chaque concept en une variable continue bornée entre 0 et 1 (variable floue). Plus un concept recevra de verbatims illustratifs, plus sa valeur sera élevée. La variable ainsi obtenue peut être assimilée à une condition qui sera plus ou moins considérée comme continue.

II. Mise en œuvre de la démarche

La démarche de transformation de textes en variable flou demande d'utiliser plusieurs logiciels. Dès lors, nous proposons de procéder en deux temps : la manipulation des données dans un logiciel d'analyse de contenu (ici Nvivo), puis la transformation de ces données en variable flou.

A. Mise en forme dans Nvivo

Établir une grille de codage alignée sur le modèle de recherche : justification théorique des variables et codage de variables émergentes le cas échéant.

Coder (compter) les verbatims par catégorie de variables. Pour ce faire, il s'agit de créer une requête dans Nvivo (ou autre logiciel) permettant d'obtenir une base de données composées en ligne les identifiants des sources de données (répondants) et en colonne des concepts codées. On obtient un tableau qui « compte » le nombre de verbatims correspondant à chaque variable pour chaque répondant. Il est important d'enregistrer ce tableau dans le logiciel d'analyse de contenu (Nvivo ?) car il permettra au chercheur de revenir sur le contenu chiffré pour extraire les séquences de textes attachés à chaque case[2].

Lancer une extraction depuis le logiciel d'analyse de contenu afin d'obtenir un fichier Excel.

B. Préparation du fichier pour FsQCA

1. Toilettage

Le fichier Excel issu de l'étape qui précède doit être nettoyé avant d'être intégré dans le logiciel d'analyse des configurations (ici FsQCA). Il est assez courant que certaines variables soient très peu complétées ou que certains individus ne soient concernés que par très peu de variables. Dans les deux cas, le tableau comporte de nombreuses valeurs nulles qui vont influencer très négativement la mise au format FsQCA.

Un premier travail de repérage des variables ou des individus à supprimer doit être réalisé. Les lignes et colonnes comportant trop de 0 doivent intégralement être supprimées. L'idée étant de repérer des lignes ou des colonnes présentant une répartition relativement complète des valeurs.

[2] Notons ici qu'il s'agit de « garder » le lien entre les valeurs numériques qui seront mobilisées dans l'analyse des configurations et le contenu sémantique attaché à ces valeurs. Dans Nvivo, en double-cliquant sur la cellule d'un tableau de requête, le logiciel affiche l'ensemble des verbatims correspondants, ainsi que leur origine. Si le chercheur souhaite identifier une série de valeurs particulières (ici une case) dans l'analyse de sa configuration finale (résultat de FsQCA), il pourra alors revenir dans le logiciel Nvivo pour fournir dans son article des verbatims qui illustrent les discours tenus par les répondants dans la configuration mise en évidence.

2. Définition des seuils

L'un des problèmes majeurs de l'utilisation des logiciels de gestion des configurations en logique floue est précisément le repérage, pour chaque variable, du seuil à partir duquel le chercheur considère que sa variable exerce ou pas une influence sur la variable expliquée. Ce seuil sera ensuite utilisé dans l'algorithme du logiciel (FsQCA) pour établir les configurations ou combinaisons de variables ayant un effet sur la variable expliquée.

Plusieurs solutions existent pour déterminer ce seuil.

La première consiste à déterminer les seuils d'influence en fonction des recherches antérieures ou du point de vue théorique. Par exemple, dans certaines études pourtant sur la propagation épidémiologique, on connait le seuil à partir duquel une certaine densité de population à risque ayant contractées certaines maladies virales (variable explicative) exercera un effet sur la propagation (variable expliquée). La détermination de ce seuil permettra dans le logiciel de déterminer quelle densité sera considérée comme risquée ou pas. Cette première approche est plus robuste bien que rarement disponible en sciences sociales, *a fortiori* dans le cadre d'études exploratoires. En effet, dans les démarches exploratoires, il est assez rare de disposer d'hypothèses sur l'effet attendu d'une condition sociale sur une autre. Si le chercheur dispose d'éléments justifiées et robustes lui permettant de fixer ce seuil, il est préférable de les mobiliser plutôt que d'utiliser la seconde solution présentée ci-dessous.

La seconde solution, la plus courante, consiste à établir en fonction de la distribution des données un seuil que l'on pourrait appeler « empirique » pour chaque variable. Cette solution est à prendre avec beaucoup de précaution dans la mesure où la détermination du seuil n'est pas en lien avec la nature de la variable, mais seulement issue de la disponibilité des données collectées dans l'échantillon. Si ce dernier est très spécifique ou atypique, le chercheur devra en tenir compte lors de l'interprétation de ses configurations en prenant soin de souligner le caractère contingent de ses résultats et ne pas tenter de les considérer comme généralisables.

Dans ce guide nous nous focaliserons sur la seconde solution. La méthode qui suit tente d'atténuer le plus possible les grandes différences constatées dans la distribution des fréquences, caractéristique des données par nature non bornées comme les verbatims.

A l'aide d'un logiciel libre (*Tosmana*[3] ou QCA[4] dans R), le chercheur peut aisément disposer d'une représentation de ses données facilitant le choix des seuils. En suivant le guide d'utilisation, il peut éditer une à une ses variables et les visualiser (fonction *Setter*) afin de repérer les points d'inflexion en fonction de la distribution des données. Pour les variables issues des *verbatims*, il n'est pas rare de constater des distributions très excentrées (Skewness > 3). Il est donc nécessaire de choisir un point d'inflexion qui permette de répartir les réponses de la manière la plus homogène possible de part et d'autre de ce point d'inflexion. Dans le cas des distributions symétriques, la moyenne pourra constituée un bon indicateur de seuil, alors

[3] https://www.tosmana.net (lien de téléchargement consulté le 18/11/15)
[4] http://www.compasss.org/software.htm#QCA-R (lien de téléchargement consulté le 18/11/15)

que dans le cas des distributions asymétriques, le chercheur pourra opter pour la médiane. Le logiciel *Tosmana* permet de repérer rapidement quelles est la valeur de ce seuil pour chaque variable. La dernière version permet en outre d'établir des groupes homogènes (clusters) calculé automatiquement.

A ce stade, le chercheur notera dans un fichier à part (dans Excel par exemple) la valeur des seuils d'inflexion pour chaque variable. Il est important également qu'il note dans le même fichier les valeurs minimum et maximum qu'il va retenir pour chaque variable. Ces valeurs sont requises pour paramétrer la fonction *Fuzzy* qui va permettre d'homogénéiser toutes les variables et les rendre utilisables dans le logiciel de configuration (FsQCA). Il est important de noter qu'il n'est pas conseillé de retenir la valeur la plus grande pour le maximum, ni la valeur la plus petite pour le minimum pour chaque variable, dans la mesure où la variable floue qui sera construite à partir de ces seuils (*min* et *max*) renverra des valeurs approchantes de 0 pour le *min*, et de 1 pour le *max* et fixera le seuil (*cross-over)* à 0,5.

3. « Floutage » des variables

Les logiciels d'analyse de configurations présentés plus haut (FsQCA, Tosmana ou QCA dans R) disposent d'une fonction permettant de transformer n'importe quelle variable en variable floue. Or, cet outil n'est efficace que pour les variables répondant aux conditions suivantes[5] :

1. le nombre de répondants est supérieur à 30 ;
2. la distribution de la variable est normale ou relativement centrée ;
3. les écarts entre les différentes valeurs de la variable sont faibles ;
4. la variable contient peu de 0.

Compte tenu de leur nature des variables numériques issues des *verbatims*, ces conditions très rarement remplies. Le néophyte sera alors tenté de renoncer à « flouter » ses variables et ne pourra pas analyser ses configurations. En effet, la plupart des fonctions de transformation des données de codage des verbatims en variables floues ne fonctionnent pas correctement. Il est toutefois conseillé de tester cette fonction de transformation au préalable dans les trois logiciels cités (FsQCA, Tosmana et QCA/R). En cas d'échec, il est encore possible de transformer ces données à l'aide du guide ci-dessous.

Nous proposons de contourner le problème en créant une macro dans Excel (tableau 1) permettant de transformer n'importe quelle variable - quelles que soient ses caractéristiques numériques, en variable floue.

La démarche est simple. Il suffit de créer une nouvelle macro dans Excel (ou dans le logiciel libre Calc) en réalisant un copier/coller du script Visual Basic fournit dans le tableau 1 ci-dessous.

[5] Notons ici que les conditions énoncées ne sont absolument pas issues d'un calcul statistique mais plutôt induites par le retour d'expérience des utilisateurs du logiciel.

Tableau 1 : Script de la fonction « fuzzy » pour Excel

```
Function FUZZ(value, lower_thresh, crossover, upper_thresh)
' FUZZ - calibrate fuzzy-sets for QCA
' This function implements the "direct" method of transforming an
   ' interval-ratio variable into a fuzzy set, as described in Ragin,
   ' Charles C. (2008) _Redesigning Social Inquiry_, Chapter 5
   ' ("Calibrating Fuzzy Sets")
' To use, enter "=FUZZ(value, lower_thresh, crossover, upper_thresh)"
   ' in a cell, parameters may be literals or cell references.  Note that
   ' the FUZZ function will NOT be available in the Function Wizard; neither
   ' can it be specified as part of a formula within the Function
   ' Wizard.  (If somebody knows how to implement this, please contact me.)
' To install as an OpenOffice/LibreOffice Calc macro:
   '
   ' In Calc, "Tools" -> "Macros" -> "Organize Macros"
   '    -> "OpenOffice (or LibreOffice) Basic"
   '
   ' In the dialog box that opens, navigate to "My Macros" -> "Standard"
   '    -> "Module1" and click "Module1" to highlight it.  Click "Edit".
   '
   ' Paste the contents of this file to the end of the window that opens
   ' (from the "Function FUZZ" statement above through "End Function", below)
   '
   ' Save the file and close it.
   '
   ' The FUZZ function should now be available in all instances of Calc (for
   ' your user).
scalar_above = 3.0/(upper_thresh - crossover)
   scalar_below = -3.0/(lower_thresh - crossover)
   scalar_at = 0
deviation = value - crossover
   if deviation < 0 then
            log_odds = deviation * scalar_below
   elseif deviation = 0 then
      log_odds = deviation * scalar_at
   else  ' deviation > 0
      log_odds = deviation * scalar_above
   endif
' I don't know how to generate a proper error
   if (upper_thresh <= crossover) or (crossover <= lower_thresh) then
            FUZZ = "#Err1"
   else
      FUZZ = exp(log_odds)/(1+(exp(log_odds)))
   endif
End Function
```

Après avoir copié le script dans l'exécutif VB d'Excel (cf. indications fournies dans le tableau 1), la fonction est opérationnelle et vous permet de calculer toutes vos variables floues. Pour ce faire, le chercheur doit alors se munir de son tableau des valeurs, minimum, maximum et seuil établi à la suite de l'analyse des distributions de ses variables. Dans la cellule résultat, la

fonction *fuzzy* (=fuzzy(variable ; min ; seuil ; max)) renvoie la valeur de la variable floue. Ce calcul se fait donc pour une seule variable et un seul individu à la fois. Après avoir donné un nom à chacune des variables floues (1ère ligne du fichier), le chercheur peut réaliser une copie automatique de sa formule vers le bas à partir de la première cellule de résultat. Il obtient ainsi les valeurs floues de sa variable pour l'ensemble des individus. Il passe ensuite à la seconde variable et reproduit cette opération pour chaque variable explicative et le cas échant pour sa variable expliquée.

Après avoir créées toutes les variables floues à la suite de celles extraites de Nvivo (il est important de les garder dans le fichier pour faciliter le repérage des *verbatims* lors de l'analyse des configurations). On obtient un fichier Excel qu'il faut convertir en fichier « .csv » pour qu'il puisse être accepté par les logiciels de configuration[6].

Notre but n'est pas ici de fournir un guide de mise en œuvre des logiciels de configurations. De très nombreux ouvrages et articles sont déjà disponibles pour couvrir ce besoin sur le site suivant : http://www.compasss.org/software.htm

III. Conclusion

Compte tenu de la grande dynamique de la communauté QCA. Le risque d'obsolescence de cette note méthodologique est important. Il nous paraît important de mettre en garde le lecteur vis à vis de ce risque. L'amélioration continue des dysfonctionnements des logiciels cités dans cette note peut rendre inutile la transformation fastidieuse des données par un script dans Excel, ajoutant ainsi une étape à la liste déjà longue de cette démarche de transformation de variable.

Notons en outre que les analyses de configuration sont souvent problématiques car dans la plupart des cas plusieurs configurations ou combinaisons de variables explicatives sont possibles pour expliquer un même phénomène. Ces différentes configurations sont parfois contradictoires et rendent délicate l'interprétation.

La méthode proposée ici permet d'approfondir l'analyse en retournant à la source des données. Pour chaque configuration, les logiciels de configurations permettent d'afficher les individus (ou cas) concernés, le chercheur peut ainsi retrouver dans Nvivo les verbatims de ces individus et fournir des interprétations aux situations étudiées. On dispose ainsi d'un dispositif qui combine les avantages des deux approches méthodologiques (qualitative et quantitative) sans perdre d'informations au cours de la réduction des données, et combine ainsi le meilleur des deux mondes.

[6] Selon les versions Excel et les systèmes d'exploitation (OSX ou Windows), il est possible que FsQCA ne reconnaisse pas le format CSV. Il est alors conseillé d'essayer plusieurs formats csv. Attention également au « nommage » des variables, celui-ci doit être court (pas plus de 8 caractères) et ne pas comporter de signes proscrits (espace, tirets, opérateur (* ; +) etc.).