TP2

Claude Alie

Introduction

L'hypothèse à l'étude dans le cadre de ce travail en est une de nature méthodologique. Nous nous intéressons à l'utilisation de dictionnaires de fréquence d'occurrences dans un sondage pour traiter de l'importance relative des thèmes abordés dans des questions ouvertes. Cette utilisation se fonde sur la présomption que les fréquences d'occurrences de mots clés reflètent l'importance relative accordée par les répondants aux thèmes abordés. Mais est-ce bien le cas? Les limites du présent travail ne permettent pas de mener des tests de validité en bonne et due forme. Nous pouvons cependant vérifier si des mots clés liés à certaines notions permettent d'opérationnaliser correctement des hypothèses invoquant ces notions. On choisira des hypothèses substantives suffisamment certaines pour que toute différence observée par rapport aux attentes soit attribuable non pas à l'hypothèse substantive elle-même, que l'on sait correcte, mais à sa mesure par fréquence d'occurrences d'un mot clé. Pour ce faire, nous utiliserons l'étude électorale canadienne de 2021 (EEC, ou CES en anglais), qui comporte des sondages pré- et post-électoraux portant sur l'élection fédérale de 2021 au Canada. Ce sondage inclut une question ouverte sur les enjeux importants en politique canadienne durant cette période selon les répondants. L'importance de ces enjeux pour les répondants est mesurée par un dictionnaire de fréquence d'occurrences de mots clés liés à plusieurs enjeux (économie, immigration, soins de santé, habitation, éducation, étique, bien-être, etc.) (voir St. Jean, 2023). La base de données générée par ce dictionnaire s'ajoute à celle du sondage proprement dit. Pour ce travail, nous examinerons deux hypothèses substantives assez claires. La première établit un lien direct entre l'égalité des droits et la notion de bien-être: plus l'égalité des droits est importante pour les répondants, plus le bien-être est un enjeu important, et plus le mot clé qui lui est associé (*welfare*) devrait apparaître souvent dans les réponses à développement. La seconde hypothèse substantive examinée ici établit un lien entre l'attitude concernant l'aide apportée aux immigrants et la notion d'immigrants elle-même: plus les répondants trouvent qu'on n'en fait pas assez ou qu'on en fait trop pour eux, plus ils considèrent qu'il s'agit d'un problème, et donc d'un enjeu important, et plus le mot clé immigration devrait se trouver dans leurs réponses sur les enjeux canadiens importants. Si ces deux hypothèses substantives se trouvent confirmées, le dictionnaire de fréquence aura joué son rôle d'instrument de mesure correctement, et sa validité comme instrument s'en trouvera renforcée.

Données et méthodes

Bases de données. L'étude électorale canadienne de 2021 (EEC, ou CES en anglais) fournit deux bases de données : la base de données principale (2021 Canadian Election Study v2.0.dta), qui inclut les données liées aux questions de sondage à choix multiples, et le dictionnaire de fréquences d'occurrences qui lui est associé, qui inclut les réponses à la question ouverte sur les enjeux les plus importants au Canada (à ce moment), à savoir la question *cps21_imp_iss*.

Nettoyage. A) *Jointure.* Pour le présent travail, ces deux bases de données, qui comportent le même nombre de données, sont tout d'abord jointes (en conformité avec l'option 1) de manière à ajouter aux colonnes de la première (une variable par question à choix multiple) celles de la seconde (une variable par mot clé reflétant un enjeu). La jonction s'effectue par la variable d'identification des réponses *cps21_ResponseId*.

B) Sélection. Les variables à l'étude pour les deux hypothèses substantives examinées sont ensuite sélectionnées, incluant :

1) Deux variables d'entiers tirées du dictionnaire de fréquence, à savoir *immigration* et *welfare*, qui indiquent la fréquence des mots clés du même nom. Ces variables agiront à titre de variables dépendantes dans les analyses de régression catégorielles.

2) Deux variables ordonnées tirées de la base de données principale, qui serviront de variables causales :

a. *pes21_equalrights* (renommée *equalrights*): *We have gone too far in pushing equal rights in this country.* Options: 1: Strongly disagree; 2: Somewhat disagree; 3: Neither agree nor disagree; 4: Somewhat agree; 5: Strongly agree; 6: Don't know / Prefer not to answer).

b. *cps21_spend_imm_min* (renommée *spend_imm_min*): *How much should the federal government spend on immigrants and minorities?* Options: 1: Spend less; 2: Spend about the same as now; 3: Spend more; 4: Don't know / Prefer not to answer).

3) Deux variables catégorielles tirées de la base de données principale, qui serviront de variables de contrôle et de critères sélection :

a. *pes21_votechoice* (renommée *votechoice*) : *Which party did you vote for?* Options: 1: Liberal Party (LP); 2: Conservative Party (CP); 3: NPD; 4 Bloc Québécois (BQ); 5: Green Party (GP); 6: People's Party (PP); 7: Another Party (AP); 8: I spoiled my vote; 9: Don't know / Prefer not to answer.

b. *pes21_province* (renommée *province*): In which province or territory are you currently living? Options: 1: Alberta (AL); 2: British Columnbia (BC); 3: Manitoba (MA); 4: New Brunswick (NB); Newfoundland and Labrador (NL); 6: Northwest Territories (NT); 8: Nunavut (NU); 9: Ontario (ON); 10: Prince Edard Island (PEI); 11: Quebec (QU); 12 Saskatchewan (SA); 13: Yukon (YU).

C) Filtres. Le niveau *Don't know / Prefer not to answer* a été éliminé des variables catégorielles, de même que *I spoiled my vote* pour la variable *votechoice*. Le niveau *Quebec* a aussi été éliminé faciliter le contrôle des variables causales liées aux particularités du Québec (langue officielle et présence du Bloc Québécois). Les données manquantes (NA) ont aussi été filtrées.

D) Regroupements et résumés statistiques. Pour l'hypothèse 1 (*equalrights* → *welfare*), les données ont été regroupées par *equalrights* pour produire un tableau de moyennes et d'écarts type avec les variables *equalrights, welfare, votechoice et province*. Pour l'hypothèse 2 (*spend_imm_min* → *immigration*), les données ont été regroupées par *spend_imm_min* et un tableau similaire été produit pour *spend_imm_min*, *immigration*, *votechoice*¸ *province*.

E) *Tests.* Les deux hypothèses à l'étude ont été testées dans une optique d'analyse causale en utilisant un modèle de régression linéaire avec les variables de contrôle *province* et *votechoice*. Des tests post hoc ont été effectués pour localiser les niveaux catégoriels affectés. Des histogrammes ont été utilisés pour visualiser les différences.

Résultats

Hypothèse 1 (equalrights → welfare). Comme le montre la Figure 1 (gauche), plus les répondants considèrent que l'égalité des droits doit être supportée, plus ils invoquent souvent la notion de bien-être (*welfare*). Ces différences sont globalement significatives ($p < 0.0001$).

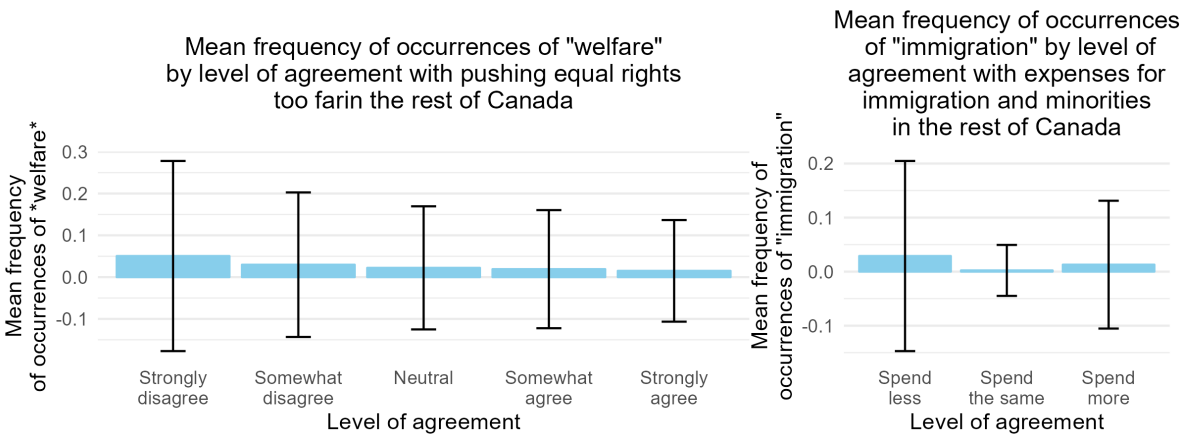

Figure 1: Terminal commands 1

Hypothèse 2 (spend_imm_min → immigration). La Figure 1 (droite) montre que lorsque les répondants sont en désaccord le niveau d'aide fourni aux immigrants et aux minorités, ils invoquent plus souvent la notion d'immigrants ($p < 0.0001$).

Conclusion. Pour les deux hypothèses investiguées, lorsque les répondants considèrent que quelque chose est important et constitue un enjeu, ils utilisent davantage de mots clés reliés

à cet enjeu. Cela supporte l'idée que le dictionnaire de fréquence d'occurrence de mots clés peut constituer un instrument de mesure approprié pour évaluer l'importance accordée à des enjeux. Cependant, la faiblesse des analyses post hoc pour localiser les différences par niveau et l'importance des écarts types (voir les moustaches de la Figure 1) rappellent que cet outil ne dispose que d'une faible puissance en comparaison des réponses à choix multiples. En effet, chaque mot clé n'est utilisé qu'occasionnellement par les participants, et d'une manière moins standardisée qu'une question à choix multiple (les mots clés peuvent avoir des résonnances différentes selon les participants). De tels dictionnaires doivent donc être utilisés avec précaution, et dans une optique essentiellement exploratoire.

Annexes

Lecture des deux bases de données brutes

Lecture de la base de données principale multi-choix *2021 Canadian Election Study v2.0.dta,* qui deviendra *canElStudy*, et de la base de données dictionnaire de fréquences *CES21_dictionarycoding_public_release_final.dta*, qui deviendra *dictCoding_pubRelease_fin_2021*.

```
canElStudy <- read_dta("../data/dataverse_files/2021 Canadian Election Study v2.0.dta")

dictCoding_pubRelease_fin_2021 <- read_dta("../data/dataverse_files/CES21_dictionarycoding
```

Création de la base de données *dat* par junction, nettoyage et filtre

La base de données *dat* contient toutes les données qui seront utilisées dans l'étude, mais ne sont pas encore traitées pour effectuer des calculs ou des graphiques. Les bases de données brutes sont d'abord fusionnées, les variables à l'étude sont sélectionnées, les données écartées sont filtrées, les variables aux noms inadéquats sont renommées, et les données manquantes sont filtrées. En plus de servir dans les analyses de régression, *dat* servira à construire les les base de données de regroupements *group_by_equalrights* et *group_by_spend_imm_min*, utilisées pour les graphiques.

```
dat <- left_join(
  canElStudy,
  dictCoding_pubRelease_fin_2021,
  by = c("cps21_ResponseId" = "cps21_ResponseId")) %>%
  select(pes21_province, pes21_votechoice2021,
         pes21_equalrights, cps21_spend_imm_min,
         immigration, welfare) %>%
```

```
  filter(pes21_province != 11) %>%
  filter(pes21_votechoice2021 != 8) %>%
  filter(pes21_votechoice2021 != 9) %>%
  filter(pes21_equalrights != 6) %>%
  filter(cps21_spend_imm_min != 4) %>%
  mutate(province = pes21_province,
         votechoice = pes21_votechoice2021,
         equalrights = pes21_equalrights,
         spend_imm_min = cps21_spend_imm_min
         ) %>%
  na.omit()
```

Création de la base de données par regroupement *group_by_equalrights* pour l'hypothèse 1

Cette base de données permettra d'effectuer des statistiques descriptives (moyennes et déviations standard) et d'effectuer les histogrammes pour vérifier l'hypothèse 1. Le regroupement s'effectue par la variable *equalrights*.

```
group_by_equalrights <- dat %>% group_by(equalrights) %>%
  summarise(
            equalrights_mean = mean(equalrights, na.rm=T),
            equalrights_sd = sd(equalrights, na.rm=T),

            welfare_mean = mean(welfare, na.rm=T),
            welfare_sd = sd(welfare, na.rm=T),

            votechoice_mean = mean(votechoice, na.rm=T),
            votechoice_sd = sd(votechoice, na.rm=T),

            province_mean = mean(province, na.rm=T),
            province_sd = sd(province, na.rm=T),
            ) %>%
  ungroup() %>%
  distinct()
```

Création de la base de données par regroupement *group_by_spend_imm_min* pour l'hypothèse 2

Cette base de données permettra d'effectuer des statistiques descriptives (moyennes et déviations standard) d'effectuer les histogrammes pour vérifier l'hypothèse 2. Le regroupement s'effectue par la variable *spend_imm_min*.

```
group_by_spend_imm_min <- dat %>% group_by(spend_imm_min) %>%
  summarise(
            spend_imm_min_mean = mean(spend_imm_min, na.rm=T),
            spend_imm_min_sd = sd(spend_imm_min, na.rm=T),

            immigration_mean = mean(immigration, na.rm=T),
            immigration_sd = sd(immigration, na.rm=T),

            votechoice_mean = mean(votechoice, na.rm=T),
            votechoice_sd = sd(votechoice, na.rm=T),

            province_mean = mean(province, na.rm=T),
            province_sd = sd(province, na.rm=T)

            ) %>%
  ungroup() %>%
  distinct()
```

Graphiques

Le code utilisé pour générer les histogrammes de la figure 1 est storé ici. Il est non visible.

Code pour l'histogramme de l'hypothèse 1

Ce code est basé sur *group_by_equalrights*.

Code pour l'histogramme de l'hypothèse 2

Ce code est basé sur *group_by_spend_imm_min*.

Régressions linéaires

Le code utilisé pour effectuer les analyses de régression est storé ici avec les résultats d'analyse. Les analyses sont effectuées à partir de la base non groupée *dat*.

Analyses de régression pour l'hypothèse 1

Les variables *province* et *votechoice* sont ajoutées à titre de variables de contrôle.

```
model <- lm(welfare ~  equalrights + province + votechoice, data = dat)
summary(model)
```

```
Call:
lm(formula = welfare ~ equalrights + province + votechoice, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.05888 -0.04137 -0.03051 -0.02072  2.95000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0436212  0.0062027   7.033 2.19e-12 ***
equalrights -0.0089793  0.0014279  -6.288 3.37e-10 ***
province     0.0010790  0.0005362   2.012   0.0442 *
votechoice   0.0018824  0.0015393   1.223   0.2214
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1784 on 8403 degrees of freedom
Multiple R-squared:  0.005294,  Adjusted R-squared:  0.004939
F-statistic: 14.91 on 3 and 8403 DF,  p-value: 1.122e-09
```

Des analyses post hoc sont effectuées.

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts
```

```
Fit: lm(formula = welfare ~ equalrights, data = dat)

Linear Hypotheses:
            Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0 -0.020735   0.005389  -3.848  0.00111 **
3 - 1 == 0 -0.031411   0.005688  -5.523  < 0.001 ***
4 - 1 == 0 -0.028284   0.005908  -4.787  < 0.001 ***
5 - 1 == 0 -0.035469   0.006798  -5.218  < 0.001 ***
3 - 2 == 0 -0.010676   0.006089  -1.753  0.39699
4 - 2 == 0 -0.007549   0.006295  -1.199  0.74867
5 - 2 == 0 -0.014735   0.007137  -2.065  0.23225
4 - 3 == 0  0.003127   0.006553   0.477  0.98922
5 - 3 == 0 -0.004058   0.007365  -0.551  0.98150
5 - 4 == 0 -0.007185   0.007536  -0.953  0.87386
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Analyses de régression pour l'hypothèse 2

Les variables *province* et *votechoice* sont ajoutées à titre de variables de contrôle.

```r
model <- lm(immigration ~  spend_imm_min + province + votechoice, data = dat)
summary(model)
```

```
Call:
lm(formula = immigration ~ spend_imm_min + province + votechoice,
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.02808 -0.01867 -0.01380 -0.00767  2.00190

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.0273143  0.0050071   5.455 5.03e-08 ***
spend_imm_min -0.0103487  0.0018306  -5.653 1.63e-08 ***
province       0.0007830  0.0003547   2.208   0.0273 *
votechoice     0.0001342  0.0010230   0.131   0.8956
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.118 on 8403 degrees of freedom
Multiple R-squared:  0.004403,  Adjusted R-squared:  0.004048
F-statistic: 12.39 on 3 and 8403 DF,  p-value: 4.412e-08
```

Des analyses post hoc sont effectuées.

```r
dat$spend_imm_min <- as.factor(dat$spend_imm_min)
model <- lm(immigration ~ spend_imm_min, data = dat)
contrasts <- glht(model, linfct = mcp(spend_imm_min = "Tukey"))
summary(contrasts)
```

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = immigration ~ spend_imm_min, data = dat)

Linear Hypotheses:
            Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0 -0.026604   0.002913  -9.133  < 0.001 ***
3 - 1 == 0 -0.015925   0.003684  -4.323  < 0.001 ***
3 - 2 == 0  0.010679   0.003456   3.090  0.00556 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```