

EEE et EEE

Essai sur l'erreur écologique et vote sur l'Espace économique européen*

Matthias Brunner et Simon Hug
Centre interfacultaire de gérontologie
et
Département de science politique
Université de Genève[†]

Papier préparé pour présentation au congrès de
l'Association Suisse de Science Politique
Balstahl, 1994

*Eva Anduiza, Danielle Bütschi, Cédric Dupont et Pascal Sciarini ont eu le courage et l'aimabilité de se pencher sur des versions antérieures et souvent embryonnaires du présent papier. Sans leurs suggestions, ce texte se trouverait dans un état encore plus déplorable. Thomas Widmer nous a dépanné en nous fournissant une information en dernière minute. Nous tenons à les remercier.

[†]59, rte de Mon-Idée; 1226 Thônex et 102, Bd Carl-Vogt; 1211 Genève 4; Suisse. Fax: + 41 22 741 81 00, email: brunnerm@ibm.unige.ch, simhug@ibm.unige.ch

1 Introduction

Les analyses de données agrégées ont connu un certain regain d'intérêt dans la littérature en science politique. Au niveau méthodologique les problèmes d'inférence écologique sont de nouveau devenu l'objet de développements intéressants (Thomsen, 1987; Berglund et Thomsen, 1990; Erbring, 1990; Claggett et van Winden, 1993, Achen et Shively, 1995). Ces nouvelles contributions montrent la rencontre malchanceuse entre la montée des enquêtes par sondage et la publication d'un article dévastateur pour l'analyse de données agrégées, à savoir la contribution de Robinson (1950). Cet article a discrédité les analyses écologiques de sorte que celles-ci ont largement disparu de la littérature.

Dans le contexte suisse, des études empiriques se basant sur des données communales (Hug, Vatter et Kummer, 1993; Ladner, 1993; Möckli, 1993; Ladner et Meuli, 1994; Hug et Kummer, 1994 et Vatter, 1994) ont été réalisées à l'occasion des récentes votations.¹ Ces travaux apportent un nouveau regard sur le comportement des citoyennes² et citoyens lors des élections. Toutefois, on note dans certaines de ces contributions un antagonisme mal placé entre les analyses écologiques et les traditionnels sondages VOX (par exemple Möckli, 1993; Vatter, 1994 et d'une manière plus générale Nef, 1989: 560). Une appréciation de l'utilité réciproque des analyses basées sur les deux types de données manque souvent.

À part ce manque de respect mutuel entre des recherches utilisant les deux types de données, nous constatons deux autres lacunes dans les travaux récents utilisant des données agrégées. D'abord, les problèmes méthodologiques inhérents à la fois aux recherches par sondage et aux analyses écologiques, donc de données agrégées, ne sont guère discutés et pris en compte. Les récents développements au niveau méthodologique (voir ci-dessus) ne trouvent pas de place dans ces travaux et les analyses en souffrent clairement. Ensuite, ces travaux récents, à part quelques comparaisons entre les deux types d'analyse (Hug et Kummer, 1994 et Vatter, 1994), ne tentent guère de combiner les deux approches.³

Ce sont ces derniers points que nous tentons de discuter dans le présent papier. En utilisant les données recueillies à l'occasion du vote sur l'Espace économique européen (EEE), nous voulons montrer les problèmes méthodologiques inhérents aux

¹Citons aussi les travaux antérieurs de Nef (1979, 1980, 1988, 1989), Joye (1984) et Meier-Dallach et Nef (1987).

²Nous remercions Danielle Bütschi pour nous avoir signalé cette omission importante dans une version antérieure du texte.

³Naturellement, ce deuxième reproche ne s'adresse guère aux papiers rédigés avant la réalisation de l'enquête VOX (Hug, Vatter et Kummer, 1993; Ladner, 1993 et Ladner et Meuli, 1994).

deux approches. En plus, nous proposons des modèles qui permettent de combiner les deux niveaux d’analyses. Par conséquent, notre démarche s’inscrit dans une perspective à la fois méthodologique et empirique. Dans la prochaine section nous posons le problème des analyses au niveau individuel. La section suivante sera consacrée au niveau agrégé et aux problèmes d’agrégation. Les modèles “multilevel” se trouvent au centre de la quatrième section, qui mène vers une conclusion mettant en perspective nos résultats.

2 Le niveau individuel

Dans l’analyse des votations en Suisse, les sondages VOX jouent un rôle prédominant.⁴ Ils fournissent régulièrement depuis 1977 des données individuelles portant sur le comportement des citoyennes et citoyens lors des votations fédérales. Cette longue série d’enquêtes,⁵ ainsi que les analyses qui en ont été faites, permettent de montrer les problèmes qui peuvent se poser pour des analyses au niveau individuel.

D’une manière générale, la prédominance d’analyses au niveau individuel s’explique largement par l’impact considérable que l’article de Robinson (1950) a eu sur la communauté scientifique (Huckfeld et Sprague, 1993: 285). Trouvant une relation négative (mesurée par un coefficient de corrélation) entre le pourcentage d’étrangers et l’illétrisme au niveau des états aux Etats-Unis, mais une relation positive sur la base de données individuelles, l’auteur conclut que:

“The purpose of this paper will have been accomplished, however, if it prevents the future computation of meaningless correlations and stimulates the study of similar problems with the use of meaningful correlations between the properties of individuals.” (Robinson, 1950: 357)

Conjugé avec une popularité grandissante des enquêtes par sondage, cet article a diminué singulièrement le recours à des données agrégées dans la littérature. Suivant la conclusion de Robinson, la majorité des chercheurs pensait que les “corrélations qui font sens” se calculent par définition au niveau individuel. Que ceci ne soit pas forcément le cas n’a pas été accepté dans la littérature pendant longtemps. Toutefois, toute une série de problèmes sont inhérents à une approche

⁴Gruner et Hertig (1983) et Kriesi (1993) rapportent des études portant sur plusieurs enquêtes VOX.

⁵Notons toutefois que seule une partie de l’ensemble des enquêtes est encore disponible. Les fichiers des votations ayant eu lieu entre 1977 et 1980 ont malheureusement été perdus.

| | participation VOX | participation réelle | vote VOX | vote réel |
|------|----------------------|-------------------------|-------------|--------------|
| oui | 89.2 | 78.3 | 50.6 | 49.7 |
| non | 10.8 | 21.7 | 48.8 | 50.3 |
| vide | - | - | 0.6 | - |
| n | 985 | | 875 | |

Table 1: Participation et vote: sondages et réalité

qui se base sur des données individuelles, recueillies à l'aide d'enquêtes par questionnaire. Dans ce qui suit, nous relevons quatre problèmes particuliers, que nous considérons comme étant les plus importants.

2.1 Problèmes de mesure

Un problème omniprésent en science sociale concerne la mesure adéquate des phénomènes auxquels le chercheur s'intéresse (Kaplan, 1964: ch.5). Dans les enquêtes par questionnaire, la question posée à l'interviewé fonctionne en tant que stimulus. La réponse fournie par l'enquêté et enregistrée par l'enquêteur produit la mesure d'un phénomène donné. Vu le processus complexe qui se déroule entre la réception du stimulus et la transcription d'une réponse (Feldman, 1991), il est évident que de multiples erreurs peuvent s'introduire. La question peut être mal formulée, rendant la tâche de l'enquêté difficile, voire impossible (Converse et Presser, 1986). Elle peut créer des malentendus ou bien produire un stimulus non voulu. L'enquêté de son côté peut refuser de réagir comme prévu à la question, etc.

D'une manière générale de multiples sources d'erreurs potentielles existent donc. Un exemple concret nous est fourni par la surestimation considérable des taux de participation aux scrutins à partir des sondages VOX. Presque systématiquement, les résultats calculés sur la base des sondages aboutissent à un pourcentage de participation plus important. Le tableau 1 montre que l'écart lors du vote sur l'EEE dépassait les dix pourcents. De même, l'écart entre le résultat du vote effectif et le pourcentage d'acceptation ne se situe que rarement dans un intervalle de confiance qu'on pourrait attribuer à des variations aléatoires d'échantillonnage. Lors du vote sur l'EEE cet écart a été contre toute attente extrêmement faible (Tableau 1).

Ces écarts constatés dans les enquêtes VOX ont au moins deux raisons distinctes. La première, que nous venons de discuter, est celle de la mesure. Concernant la participation, la question "Avez vous participé . . ." fournit bien le stimulus à l'interviewé, mais sa réponse peut être modifiée par certaines attentes. La participation aux scrutins est souvent une norme sociale, et s'abstenir n'est pas forcément un acte dont on se vante en public. D'une manière consciente ou inconsciente, la

réponse au stimulus peut être biaisée. La deuxième raison possible des écarts se retrouve dans les problèmes d'échantillonnage, que nous discuterons plus en détails ci-dessous.

Le sens de certaines questions peuvent aussi changer dans le temps. Au début des années 50 une question fréquemment posée à des Américains était la suivante: "Do you like to go to gay parties?" Cette question fournissait pendant une longue période des informations importantes sur les loisirs des individus. Au début des années 70 et surtout dans les années 80 la perception des termes utilisés dans cette question a singulièrement changé. Ceci illustre qu'une mesure peut perdre sa validité (Kaplan, 1964: 198), puisqu'elle cesse de mesurer ce qu'elle devrait mesurer.

Un autre problème tout aussi important et celui de la "reliability" (Kaplan, 1964: 200). Toute mesure devrait fournir la même réponse, si elle est appliquée d'une manière répétée. Des questions difficiles à comprendre sont souvent peu fiables, puisque la réponse fournie par des individus n'est que faiblement dépendante du stimulus, donc de la question.

La réaction de tout défenseur de l'approche individuel face à ces problèmes sera d'avancer que les mêmes difficultés se manifestent avec l'utilisation de données agrégées. Ceci est largement vrai, puisque une grande partie des données statistiques, fournies par des recensements, sont recoltées à l'aide de questionnaires. Toutefois, pour certains phénomènes, les mesures au niveau agrégé sont moins problématique qu'au niveau individuel. Déjà par la simple institution du vote secret, les pourcentages d'acceptation dans une commune sont entachés d'une erreur considérablement plus faible que ce qu'un sondage par questionnaire pourrait donner. Mais cet exemple n'est guère seul. Un grand nombre d'informations sensibles, par exemple les fortunes des individus, la possession d'une voiture, la structure de l'âge, peut être obtenue d'une manière plus fiable au niveau agrégé qu'au niveau individuel.

Ces exemples montrent que des problèmes de mesure ne désavantagent pas forcément une approche vis-à-vis de l'autre. Les deux types de données sont entachés d'erreurs de mesures. Certaines erreurs sont plus importantes au niveau agrégé, d'autres, au contraire, le sont plus au niveau individuel. L'argument pourrait être fait que l'intérêt de tout chercheur utilisant des données individuelles se porte à des comportements de personnes. Donc, le fait de savoir que pour certains phénomènes des mesures plus fiables et précises sont disponibles au niveau agrégé ne semble peu utile. Comme nous tentons de montrer par la suite, cet argument n'est que

partiellement convaincant.

2.2 Problèmes de représentativité

Nous avons vu dans la section précédente que des erreurs de mesure peuvent entacher les données de sondage par questionnaire. Une autre source d'erreur provient de la représentativité de l'échantillon à partir duquel on veut inférer au niveau de la population. Dans ce texte, nous nous bornerons à distinguer deux problèmes liés à l'échantillonnage: d'abord celui des échantillons non-aléatoires puis celui de la représentativité de la stratification utilisée.

2.2.1 L'échantillonnage non-aléatoire

Dans le cas idéal, l'échantillon provient du tirage aléatoire d'un certain nombre d'individus sur une liste de population.⁶ Chaque individu dans la population a une probabilité identique et calculable d'appartenir à l'échantillon. Le problème de représentativité provient alors essentiellement de la fluctuation aléatoire de l'échantillonnage. Selon le théorème de la limite centrale (Central Limit Theorem), cette fluctuation suit une distribution normale. Il est alors possible d'estimer le biais d'échantillonnage à partir de l'écart-type de la moyenne estimée (Mohr, 1993:110-119).

Malheureusement, en Suisse, une grande partie des données sur le comportement politique individuel provient d'enquêtes qui utilisent une méthode de sélection non-aléatoire. Les enquêtes VOX, par exemple, font recours à la méthode dite des quotas. Cette technique est différente de la méthode aléatoire (random sampling) en ce qu'elle laisse aux enquêteurs le libre choix des individus interrogés à condition de respecter certains quotas d'âge et de sexe.⁷

Si un tel procédé présente l'avantage évident d'être moins coûteux, il pose cependant un problème supplémentaire au niveau de la représentativité. Avec la méthode aléatoire, il est possible d'estimer le biais d'échantillonnage. La méthode des quotas interdit une telle procédure parce que l'indétermination du processus de sélection des individus interdit toute connaissance de la distribution de l'échantillonnage.⁸ Le

⁶Dans la pratique, ce tirage aléatoire s'effectue généralement en deux temps. On sélectionne d'abord aléatoirement un certain nombre d'unités géographiques (en Suisse par exemple des communes) au sein desquelles on va, dans un second temps, consulter les registres de population.

⁷Une enquête par quotas procède en général comme une enquête aléatoire en ce qui concerne la sélection des unités géographiques, mais laisse à l'enquêteur le libre choix des individus à l'intérieur des quotas.

⁸La différence entre les deux méthodes apparaît clairement au niveau du traitement des refus de répondre. Dans la méthode aléatoire, les individus sélectionnés seront recontactés plusieurs fois et, dans certains cas, remplacés par des individus qui leur ressemblent. Non seulement on garde le compte des refus, mais on tente de les limiter au maximum. Dans une enquête par quotas, il est impossible de déterminer le nombre de refus pour une réponse.

débat sur les inconvénients de l'échantillonnage par quotas est ancien:

“The issue of quota versus probability sampling has been a matter of controversy for many years. Some experts hold the quota method to be so unreliable and prone to bias as to be almost worthless; others think that, although it is obviously less sound theoretically than probability sampling, it can be used safely on some subjects; still others believe that with adequate safeguards, quota sampling can be made highly reliable and that the extra cost of probability sampling is not worth while. In general, statisticians have criticized the method for its theoretical weakness, while market and opinion researchers have defended it for its cheapness and administrative convenience.” (Moser and Kalton, 1979[1971]:127)

Diverses méthodes peuvent être utilisées pour contourner -imparfaitement- le problème.⁹ Sans entrer dans les détails de ces procédures, nous nous bornerons ici à distinguer les cas dans lesquels la méthode des quotas pose un problème sérieux. Dans cette optique, deux situations peuvent être distinguées. Dans la première, la technique de sélection de l'échantillon, bien que non-aléatoire, est indépendante de la variable considérée; dans la seconde elle est fortement corrélée avec celle-ci.

Pour illustrer le cas de l'indépendance, imaginons une enquête qui s'intéresse au pourcentage de personnes myopes dans la population. Si les données proviennent d'un échantillonnage par quotas, on peut raisonnablement faire l'hypothèse que la sélection des individus est indépendante de la myopie éventuelle de la personne interrogée. Il est en effet difficile d'admettre que la propension à répondre est influencée par la qualité de la vision. Dès lors, les résultats de la méthode des quotas seront vraisemblablement proches de ceux qui auraient été obtenus à l'aide d'une technique aléatoire.

Inversement, si la variable dépendante est corrélée avec le processus de sélection, on peut s'attendre à des résultats fortement biaisés. Les enquêtes VOX connaissent ce problème si, par exemple, la variable considérée est la participation aux élections. On peut penser, en effet, qu'une personne qui n'est pas intéressée par la politique a une probabilité plus faible de répondre à un questionnaire portant sur la politique, qu'une autre qui s'y intéresse. La recherche montre aussi que cette personne non-intéressée aura une probabilité plus faible de se rendre aux urnes. Ces enquêtes

⁹Moser et Kalton (1979:132) proposent d'estimer cette dernière en répétant l'échantillonnage par quotas dans le temps (replicated sampling).

| Année | 1971 | 1975 | 1987 | 1991 |
|-----------------------|-----------|-----------|--------|--------|
| Echantillonnage | Aléatoire | Aléatoire | Quotas | Quotas |
| Participation estimée | 60.8 % | 53.6 % | 64.5 % | 63.7 % |
| Participation réelle | 56.9 % | 52.4 % | 46.5 % | 46.0 % |
| Erreur | 3.9 % | 1.2 % | 18.0 % | 17.7 % |
| n | 1887 | 1379 | 998 | 998 |

Table 2: Comparaison des taux de participation réels et estimés

vont donc surestimer de manière substantielle le taux de participation. La participation aux élections du Conseil National telle qu'elle ressort de quatre enquêtes réalisées en 1971, 1975, 1987 et 1991 illustre parfaitement le phénomène. Les deux premières enquêtes ont été réalisées selon la méthode aléatoire; les deux autres selon la méthode des quotas. Le tableau 2¹⁰ montre que les différences sont frappantes. Le biais d'échantillonnage est de moins de 5 % pour les enquêtes aléatoires et de près de 20 % dans le cas de l'échantillonnage par quotas.

Pour illustrer les conséquences de l'échantillonnage non-aléatoire de manière convaincante, il faudrait disposer de deux enquêtes réalisées au même moment selon les deux techniques différentes. Moser et Stuart (1953:352¹¹) ont mené une telle comparaison. Leurs résultats sont contrastés: pour certaines variables, les différences sont minimales; pour d'autres, comme par exemple l'occupation professionnelle, elles sont importantes. Il semble que c'est en effet la covariation entre le processus de sélection et la variable considérée qui détermine l'étendue du biais.

2.2.2 La stratification non-représentative

Dans de nombreuses enquêtes, l'échantillon est stratifié selon divers critères, tels l'âge, le sexe, la région ou l'occupation professionnelle. En général la proportion d'individus dans chaque quota dépend de l'importance de la catégorie dans la population totale. Dans certains cas, ces proportions peuvent cependant s'écarter substantiellement des chiffres de la population. C'est par exemple le cas des enquêtes VOX par rapport aux régions linguistiques. La partie romande du pays est volontairement surestimée pour être représentative, non au niveau de la population suisse, mais à celui de la romandie. De même, on interroge dans ces enquêtes davantage d'hommes de plus de 65 ans que de femmes alors que dans la population, les femmes de cette classe d'âge sont en nette majorité (Tableau 2). Ainsi, lorsqu'on tentera

¹⁰Les chiffres de 1971 et 1975 proviennent de deux enquêtes réalisées par le Département de science politique de l'Université de Genève auprès d'échantillons représentatifs. Les chiffres de 1987 et 1991 sont ceux des deux enquêtes VOX réalisées après les élections du Conseil National. S'il est clair que la taille de l'échantillon joue ici un rôle quant au biais de l'échantillonnage, celle-ci ne saurait cependant expliquer qu'une infime partie de l'erreur d'estimation.

¹¹Cités dans Moser et Kalton (1979:131).

| Pourcentages et différences | | | | | | | | | | | | |
|-----------------------------|------------------------|------|-----------------------|------|------------------|------|------------------|------|------------------|------|------------------|------|
| Classe d'âge | Recensement 1990 (OFS) | | Enquête VOX sur l'EEE | | Enquête VOX n 46 | | Enquête VOX n 44 | | Enquête VOX n 42 | | Enquête VOX n 41 | |
| | H | F | H | F | H | F | H | F | H | F | H | F |
| 18-29 ans | 11.6 | 11.6 | 9.3 | 10.5 | 10.9 | 11.5 | 11.0 | 11.7 | 12.6 | 10.9 | 11.6 | 10.9 |
| 30-64 ans | 26.6 | 29.4 | 32.6 | 30.7 | 28.6 | 29.8 | 30.2 | 30.1 | 29.1 | 30.5 | 29.2 | 30.2 |
| 65-74 ans | 4.8 | 6.2 | 5.3 | 4.3 | 6.9 | 3.3 | 6.3 | 3.1 | 6.3 | 3.1 | 6.0 | 4.3 |
| 75 ans et + | 3.4 | 6.3 | 4.5 | 2.8 | 5.1 | 3.9 | 4.0 | 3.6 | 3.7 | 3.9 | 4.4 | 3.5 |
| Total | 46.5 | 53.6 | 51.7 | 48.3 | 51.5 | 48.5 | 51.5 | 48.5 | 51.7 | 48.3 | 51.2 | 48.8 |
| n | | | | | | | | | | | | |

Table 3: Pourcentage des citoyens par age et sexe selon le recensement, la VOX sur l'EEE et les enquêtes VOX 1986-1993

par exemple de déterminer l'influence du veuvage sur la participation aux élections, cet effet sera clairement sousestimé parce que le nombre de veufs est nettement plus bas dans l'échantillon.¹² Ce problème de représentativité de la stratification est trivial en soi, mais trop souvent ignoré dans la recherche.

Le tableau 3¹³ illustre le problème dans le cas de quelques enquêtes VOX, mais cette fois pour l'âge et le sexe. Il représente les pourcentages d'hommes et de femmes par sexe et classe d'âge parmi les citoyens qui ont le droit de vote. La distribution des échantillons en fonction des variables "sexe" et "âge" n'est pas représentative de la population. On voit d'abord que les femmes sont largement sous-représentées dans les VOX. Il apparaît de même que la répartition de l'échantillon au sein des classes d'âge s'éloigne aussi notablement de celle qui ressort du recensement, principalement en ce qui concerne la classe d'âge la plus jeune et les femmes âgées de plus de 50 ans. Voyons maintenant, dans le tableau 4,¹⁴ le taux de participation par classe d'âge tel qu'il ressort des enquêtes VOX afin de comprendre l'effet de la non-représentativité des quotas. Deux constats principaux s'imposent. D'abord, les différences de participation selon l'âge sont minimales dans le cas de la votation sur l'espace économique européen et substantielles dans les autres VOX. Ensuite, on observe que les classes d'âges qui participent le moins sont sousreprésentées dans les VOX (Tableau 3). Les trois dernières lignes montrent respectivement les taux de

¹²Etant entendu, bien sûr, que la plus grande partie des personnes veuves sont des femmes âgées de plus de 65 ans.

¹³Les chiffres des deux colonnes du recensement 1990 représentent le % de la case considérée au sein des individus suisses de 20 ans et plus. Ainsi, les femmes de 75 ans et plus représentent 6.57% des citoyens suisses de 20 ans et plus. Les chiffres des quatre colonnes restantes représentent la différence entre le % de la case parmi l'ensemble des personnes interrogées dans l'enquête considérée et le % de la case selon le recensement. Ainsi, il y a 2.93% (6.57% - 3,64%) de femmes de 75 ans et plus parmi les citoyens de 20 ans et plus interrogés lors de la VOX sur l'EEE.

¹⁴Les chiffres des deux colonnes du recensement 1990 représentent le % de la case considérée au sein des individus suisses de 20 ans et plus. Ainsi, les femmes de 75 ans et plus représentent 6.57% des citoyens suisses de 20 ans et plus. Les chiffres des quatre colonnes restantes représentent la différence entre le % de la case parmi l'ensemble des personnes interrogées dans l'enquête considérée et le % de la case selon le recensement. Ainsi, il y a 2.93% (6.57% - 3,64%) de femmes de 75 ans et plus parmi les citoyens de 20 ans et plus interrogés lors de la VOX sur l'EEE.

| Taux de participation | | | | | | | | | | |
|-----------------------|-----------------------|------|------------------|------|------------------|------|------------------|------|------------------|------|
| Classe d'âge | Enquête VOX sur l'EEE | | Enquête VOX n 46 | | Enquête VOX n 44 | | Enquête VOX n 42 | | Enquête VOX n 41 | |
| | H | F | H | F | H | F | H | F | H | F |
| 18-29 ans | 82.6 | 86.4 | 50.0 | 41.2 | 46.7 | 40.4 | 41.5 | 27.1 | 37.6 | 25.0 |
| 30-64 ans | 89.7 | 92.0 | 68.8 | 66.2 | 64.8 | 56.5 | 54.4 | 51.8 | 50.5 | 39.8 |
| 65-74 ans | 90.0 | 87.5 | 81.4 | 42.4 | 78.0 | 57.1 | 72.1 | 50.0 | 65.6 | 27.9 |
| 75 ans et + | 90.9 | 82.1 | 71.2 | 62.5 | 72.2 | 35.3 | 62.2 | 35.1 | 63.6 | 41.7 |
| Total VOX | 89.2 | | 62.8 | | 57.1 | | 49.5 | | 43.2 | |
| Total Réel | 78.4 | | 45.7 | | 38.6 | | 32.7 | | 30.6 | |
| Total pond. | 88.5 | | 61.2 | | 55.9 | | 48.5 | | 42.1 | |
| n | | | | | | | | | | |

Table 4: Le taux de participation par classe d'âge dans la VOX sur l'EEE et les enquêtes VOX 1986-1993

participation dans la VOX, en réalité, ainsi que lorsqu'on applique les proportions du recensement à la VOX. On observe que la correction apportée par la pondération est minime. Deux facteurs expliquent ce phénomène: d'abord, nous l'avons vu, le biais d'estimation provient principalement de la technique d'échantillonnage non-aléatoire; ensuite, la correction apportée ici n'est pas parfaite, tant parce qu'il faudrait utiliser la même stratification en termes de classes d'âge que celle des enquêtes, qu'à cause du fait que nous ne corrigeons pas ici par l'ensemble des facteurs de stratification utilisés dans les VOX (région linguistique, villes-campagnes, ...).

2.2.3 La pondération comme remède aux problèmes d'échantillonnage

Nous l'avons vu, la pondération peut permettre de résoudre une partie des problèmes liés à la non-représentativité de la stratification de l'échantillon. On pourrait aussi utiliser la même technique pour corriger le biais d'échantillonnage lié à la technique des quotas. Le principe serait alors de comparer la composition de l'échantillon à celle de la population afin de déterminer les caractéristiques socio-démographiques qui sont clairement sous- ou surreprésentées dans l'échantillon. La même technique peut aussi s'appliquer au taux de participation réel. Quels sont les avantages qu'on peut attendre d'une telle procédure?

Si le but recherché est la prédiction correcte de paramètres de la population, on peut espérer en tirer profit, mais de gros efforts devront être investis. Si au contraire, les fins de l'analyse ne sont qu'explicatives, ce procédé est inutile. En effet, la pondération n'ajoute pas d'information aux données individuelles, elle ne fait que modifier le poids relatif de certains groupes.

Reprenons l'exemple de l'effet de l'âge et du sexe sur la participation dans les enquêtes du tableau 3. La pondération n'est utile que si on effectue une analyse bivariée sur l'un des deux facteurs. La courbe de participation selon l'âge dans un

fichier non pondéré différerait sensiblement de celle d'un fichier pondéré parce que la proportion d'homme et de femmes n'est pas correcte, surtout dans les classes d'âge les plus vieilles. Si on dessine au contraire la courbe de chaque sexe selon l'âge dans ces deux fichiers, nous n'observerons aucune différence. Cet exemple est généralisable pour un nombre quelconque de facteurs.

Ainsi, un modèle de régression multiple bien spécifié¹⁵ ne nécessite en aucune manière une pondération des variables indépendantes. On peut même ajouter que si l'objectif est la prédiction, il est bien plus facile d'appliquer les résultats de cette régression aux données de la population plutôt que de pondérer les données du sondage.¹⁶

Enfin, il convient de rappeler ici que que la non-représentativité par rapport à une variable particulière ne pose un problème que lorsque cette variable est corrélée avec la variable dépendante considérée. Si la décision par rapport à un référendum s'explique exactement de la même manière en suisse allemande et en romandie, un échantillon composé d'un trop grand nombre de romands permettra aussi bien qu'un autre d'expliquer ou de prédire cette décision.

2.3 Problèmes de spécification

Tandis que les problèmes d'échantillonnage sont posent essentiellement au niveau technique, un autre ensemble de problèmes sont d'ordre plutôt théorique: Les problèmes de spécification peuvent apparaître de deux manières principales. D'abord, le modèle peut ne pas comprendre toutes les variables importantes. Plus précisément on peut avoir inclus une variable irrelevante ou au contraire omis une variable déterminante. Ensuite, la forme fonctionnelle postulée de la relation entre la variable dépendante et les indépendantes peut être fausse.

2.3.1 Omission d'une variable déterminante

L'inclusion d'une variable irrelevante dans un modèle de régression multiple n'a qu'une influence très limitée sur l'estimation des autres coefficients.¹⁷ Nous n'appro-

¹⁵Nous adoptons ici la définition de Achen et Shively (1995), qui considèrent qu'un modèle est bien spécifié, lorsque les coefficients estimés selon la méthode des moindres carrés sont sans biais. Dans notre contexte, ceci veut simplement dire que les variables omises qui influencent la variable dépendante ne sont pas corrélées avec les variables incluses dans le modèle.

¹⁶Le cas de la pondération par la variable dépendante mérite cependant un commentaire particulier. L'utilisation de cette technique dans le cas des enquêtes du tableau 4 ne produit que des différences minimes (variables indépendantes= sexe, âge, seul, intérêt pour la politique et interaction sexe-âge).

¹⁷La conséquence unique est l'augmentation des erreurs standards des autres coefficients (Berry and Feldman, 1993:168-170).

| | b | erreur standard | b | erreur standard | b | erreur standard | b | erreur standard |
|----------------------|--------|--------------------|-------|--------------------|--------|--------------------|--------|--------------------|
| ville | 0.488 | 0.212 | .204 | .225 | 0.192 | 0.229 | 0.251 | 0.227 |
| romande | | | | | 0.815 | 0.275 | 0.850 | 0.274 |
| éducation élevée | | | .835 | .247 | 0.918 | 0.251 | 1.052 | 0.250 |
| gauche (échelle) | | | 1.172 | .253 | 1.131 | 0.256 | | |
| (identification PSS) | | | | | | | 0.999 | 0.244 |
| constante | -0.134 | 0.173 | -.469 | .187 | -0.633 | 0.198 | -0.699 | 0.202 |
| -2 log-likelihood | | 557.942 | | 518.966 | | 513.081 | | 509.775 |
| n | | 409 | | 409 | | 409 | | 409 |

Table 5: Quatre modèles d’explication du vote sur l’EEE (regression logistique)

fondirons donc pas ce cas. La deuxième possibilité concerne l’omission d’une variable déterminante du modèle. Cette situation a, au contraire des conséquences plus sérieuses que nous allons tenter d’illustrer ici. L’exemple suivant (Tableau 5), inspiré des analyses de Kriesi et al (1993) sur le vote sur l’EEE, montre bien le problème.¹⁸ Au lendemain du vote, un nombre important de commentateurs ont souligné que ce n’étaient pas seulement les cantons romands qui ont accepté l’EEE, mais aussi les grandes villes alémaniques. Donc, on pourra supposer que le fait d’habiter dans une ville augmente la propensité de voter pour l’objet en question. Ceci est effectivement le cas si l’on estime un modèle bivarié. Parmi la totalité des personnes pour lesquelles on dispose de l’information nécessaire, la probabilité d’acceptation se situait à 0.52. Pour les personnes habitant la campagne cette probabilité tombe à 0.49, mais passe à 0.58 pour les citadins.¹⁹ Donc, le fait d’habiter en ville semble avoir un impact considérable sur le vote pour l’EEE.

Ayant vu les résultats de cette estimation, tout chercheur devrait se poser la question si le modèle est bien spécifié. Plus particulièrement, ils s’agit de savoir si aucune variable ayant un impact sur la variable dépendante n’a été omise, tout en étant corrélée avec une variable incluse dans le modèle. Pour un modèle de cette simplicité et d’un tel manque de fondement théorique, les chances sont grandes que le modèle soit mal spécifié. Ainsi, un modèle plus complet, qui tient aussi compte de la région linguistique, de l’éducation, de l’orientation idéologique de l’individu interrogé pourra s’approcher davantage d’une bonne spécification. Les changements qui interviennent suite à cette respcification du modèle sont importants au niveau substantiel. En effet, sous contrôle des autres variables, le coefficient estimé pour les

¹⁸Nous suivons dans cet exemple les recodifications que ces auteurs ont adoptées. Toutefois, nous éliminons les personnes qui n’arrivent pas à se situer sur l’échelle gauche-droite. Cette dernière est utilisée pour déterminer si un individu est de gauche ou de droite.

¹⁹Puisque le modèle est basé sur une variable dépendante dichotomique, une regression logistique a été estimée. Les coefficients estimés représentent dans ce cas le changement induit dans les “log-odds.” Mais pour l’interprétation il est préférable d’exprimer les changements induit dans la probabilité de voter pour l’EEE. Le calcul de ces changements est très simple et la formule peut être trouvée dans tout manuel de statistique.

citadins diminue considérablement. En moyenne, le modèle prédit une probabilité d'acceptation pour les personnes habitant à la campagne de 0.52. Cette probabilité monte seulement à 0.53 pour les citadins.²⁰ Il s'avère donc que sous le contrôle d'autres variables, l'effet d'habiter en ville disparaît presque complètement.²¹ Par conséquent, même au niveau individuel, le chercheur doit être attentif à des erreurs de spécification. L'exemple montre aussi que l'ommission d'une variable n'entraîne pas forcément des biais. Lorsque la variable de la région linguistique n'est pas utilisée pour expliquer le vote, les changements provoqués dans les coefficients des autres variables sont minimes. Ceci malgré le fait que la variable "romande" a un impact considérable sur le vote. L'absence de biais montre tout simplement que la région linguistique, tout en étant liée avec la variable dépendante, n'est que faiblement corrélée avec les autres variables indépendantes. L'ommission d'une telle variable n'introduit pas de biais dans les coefficients estimés.

Dans tous les autres cas des problèmes peuvent se poser. L'impact de la variable omise peut d'abord être incorrectement attribué à une autre. Si on étudie par exemple la participation au vote sur l'EEE selon le canton, on pourra être étonné du résultat de Schaffouse qui obtient une participation de 85.5% contre 78.3% pour le reste de la Suisse.²² Une analyse qui se bornerait à dire que la culture politique de ce canton semble particulière commettrait une erreur particulièrement grossière en omettant la variable déterminante: l'obligation du vote. Un Schaffousois s'expose en effet à une amende s'il ne se rend pas aux urnes. On pourrait parler dans ce cas de l'omission d'un effet contextuel.

2.3.2 La forme fonctionnelle des relations

Dans une régression linéaire, on postule la linéarité et l'additivité des relations entre les variables indépendantes et la dépendante. La linéarité renvoie à un effet constant sur la variable dépendante quel que soit le niveau de la variable indépendante

²⁰Ces probabilités sont calculées en tenant les autres variables constantes à leur moyenne respective.

²¹Le clivage urbain-rural pose un problème supplémentaire. En prenant la classification adoptée par les enquêtes VOX on se retrouve avec trois catégories: grandes villes (Zurich, Genève, Berne, Bâle et Lausanne), villes moyennes et communes rurales. En introduisant cette catégorisation comme variable indépendante en la transformant en deux variables dichotomiques, l'effet du milieu urbain s'avère plus complexe. Plus particulièrement, en prenant les communes rurales comme catégorie de référence, on s'aperçoit que les petites villes ont voté plus fortement contre l'EEE. En revanche, les habitants des grandes villes ont voté sensiblement plus pour l'EEE que les habitants de communes rurales. Les coefficients estimés pour un tel modèle sont les suivants: Grande ville .45 (.16), Ville moyenne -.32 (.20), Education élevée 1.57 (.20), romande .90 (.18), Constante -.70 (.13), n ???, erreur standard de la régression. Ces résultats illustrent de nouveau un problème important de la spécification, à savoir la linéarité d'une relation.

²²Notons toutefois, que lors de la votation sur l'EEE le canton de Zoug a connu un taux de participation plus important que le canton de Schaffouse.

(Berry and Feldman, 1993:201). On peut montrer que ce postulat est violé dans le cas d'une variable dépendante dichotomique (Aldrich and Nelson, 1984). Un traitement exhaustif du postulat de linéarité dépasserait cependant les objectifs de ce texte.²³

L'additivité renvoie à un effet constant quel que soit le niveau des autres variables dépendantes. Si l'effet d'une variable dépendante change selon les niveaux d'une autre variable, on parle d'une interaction. Ainsi, un modèle bien spécifié doit non seulement comprendre toutes les variables déterminantes, mais aussi tous les effets d'interaction significatifs. A titre d'exemple, nous prendrons ici l'impact de l'identification partisane sur le vote pour l'EEE.

Dans le tableau 6,²⁴ trois modèles de complexité croissante sont testés. Dans le premier modèle nous ne tenons compte que de l'identification partisane. La catégorie de référence est constituée des personnes qui déclarent qu'elles se sentent proches du parti Démocrate-Chrétien. Dans le tableau, on observe par exemple que les sympathisants du parti Socialiste ont une propensité plus grande à voter oui par rapport aux individus de la catégorie de référence. Ceux de l'Alliance des Indépendants, du parti évangélique populaire, du parti libéral, des formations d'extrême droite et ceux qui s'identifient à un autre parti ou refusent de répondre ont au contraire davantage tendance à voter non.

Comme on sait que les sections romandes de certains partis ont parfois donné un mot d'ordre différent de celui des alémaniques (Kriesi et al, 1993), il semble judicieux de tenir compte de la région linguistique. Lorsqu'on contrôle par l'appartenance à la romandie, les coefficients changent cependant relativement peu. En effet, ce contrôle repose sur le postulat que les différences entre les régions linguistiques sont les mêmes pour toutes les préférences partisans. Les différences entre les sympathisants du PDC et les autres restent donc les mêmes en suisse alémanique et en suisse romande. Un tel postulat est clairement abusif puisqu'on s'attend justement à des différences.

Ainsi, dans le troisième modèle, nous introduisons une interaction entre l'identifi-

²³Une illustration graphique de ce problème est présentée dans la figure 1 (ci-dessous), qui met en relation l'âge avec l'acceptation du traité EEE. Clairement, la relation entre les deux variables n'est pas linéaire.

²⁴Les coefficients B pour la suisse romande ont été obtenus en additionnant les coefficients d'interaction aux coefficients des partis. Ils représentent donc l'écart estimé entre les sympathisants romands du parti considéré et les sympathisants romands du PDC (Hardy, 1993:104). Pour calculer l'écart estimé entre les sympathisants romands considérés et les sympathisants alémaniques du PDC, il convient d'ajouter 2.00 qui est la valeur du coefficient de ROMANDE. Les tests de significativité de cette colonne nous permettent de savoir si la différence entre les sympathisants du PS et les individus considérés est significativement différente entre les régions linguistiques.

| Variable | Modèle A | | Modèle B | | Modèle C | | |
|---|----------|------|----------|------|----------|-------|--------|
| | b | e.s. | b | e.s. | b(alem) | e.s. | b(rom) |
| Identification partisane (catégorie de référence: PDC) | | | | | | | |
| Aucune | -.20 | .25 | -.21 | .26 | -.53 | .28 | 1.62 |
| PDT/POP | -.88 | 1.25 | -1.74 | 1.31 | -6.33 | 22.24 | -0.40 |
| PS | .62 | .30 | .57 | .31 | .41 | .34 | 1.21 |
| PES/AV | -.06 | .42 | -.05 | .44 | -.44 | .47 | 5.80 |
| ADI/PEP | -1.37 | .62 | -1.22 | .63 | -1.23 | .63 | -6.60 |
| PRD | .30 | .32 | .30 | .33 | .32 | .35 | .22 |
| UDC | -.85 | .38 | -.75 | .39 | -.73 | .40 | -1.79 |
| PLS | -1.06 | .58 | -1.34 | .62 | -1.74 | .82 | 0.00 |
| Vig/DS/PA | -1.98 | .80 | -2.33 | .83 | -2.33 | 1.09 | 1.66 |
| autre/NR | -1.19 | .50 | -1.21 | .52 | -1.98 | .67 | 5.79 |
| romande | — | — | 1.54 | .21 | 2.00 | 3.16 | |
| constante | -.42 | .16 | -.81 | .18 | -1.19 | 2.03 | |
| n [à compléter] | -.27 | .12 | -.70 | .13 | .08 | .15 | |
| erreurs standards | -.27 | .12 | -.70 | .13 | .08 | .15 | |

Table 6: L'impact de l'identification partisane sur le vote pour l'EEE

cation partisane et l'appartenance à la romandie. L'effet de l'identification partisane peut maintenant varier d'une région linguistique à l'autre. Comme l'interprétation des coefficients devient nettement plus compliquées lorsqu'on introduit une interaction, nous avons simplifié la présentation des coefficients en calculant les différences, en romandie par rapport aux sympathisants du PDC.

Dans ce dernier modèle, on voit que les personnes sans identification partisane se comportent très différemment selon la région linguistique. Si en suisse allemande elles semblent voter plus négativement que les sympathisants du PDC, il apparaît qu'elles ont le comportement inverse en suisse romande. Au niveau des autres identifications, aucune différence linguistique n'est significative. C'est donc seulement en considérant cette interaction qu'il nous a été possible de mettre à jour cet effet différentiel qui avait déjà été mis en avant par Kriesi (1993:xxx à vérifier) pour le vote des abstentionnistes.

2.4 Inadéquation de données individuelles

Les trois points précédents semblent partager au moins un aspect, à savoir que le problème peut être résolu relativement facilement: soit en proposant de meilleures mesures, soit en tirant des échantillons avec soin, soit en proposant des modèles bien spécifiés. Ce point est certes vrai, mais souvent, comme le montraient les derniers exemples de problèmes de spécification, le contexte ou bien des variables au niveau agrégé ne peuvent pas être tenu à l'écart. Ceci est encore plus frappant pour les exemples que nous présenterons ici. Ces derniers montrent que malgré une problématique qui semble être purement individuelle, les données à ce niveau d'analyse sont inadéquates.

L'analyse de Longchamp (Kriesi et al, 1993: 16ss), qui s'intéresse à l'impact des médias sur le vote pour l'EEE, est un exemple typique. La couverture par les médias de l'objet du vote ne peut varier qu'à travers le temps, et dans une moindre mesure dans l'espace. Du coup elle ne varie que d'un groupe d'individu à l'autre. Ces groupes sont définis par l'appartenance à des unités temporelles ou spatiales. Si l'on s'intéresse à l'impact des médias sur le vote, cette faible variation entre les unités, risque fortement d'être noyée dans des variations plus considérables au niveau individuel. Plus précisément, pour l'exemple de l'étude de Longchamp, les intentions de vote des individus varient considérablement d'une vague d'enquête à l'autre. Un part important de cette variation est dûe au hasard ou est expliquée par certains facteurs spécifiques. Mais ces derniers, qui varient d'un individu à l'autre, ont beaucoup plus de chance à covarier avec la variable dépendante qu'une variable qui reste constante à l'intérieur des groupes d'individus.²⁵ Donc, ici un problème considérable se pose, qui, à la limite, ne trouve de solution que dans une agrégation des données individuelles au même niveau sur lequel on dispose des mesure portant sur les médias.²⁶ C'est cette stratégie de recherche qui est empruntée par Longchamp. Il tente d'expliquer les variations dans les opinions favorables à l'EEE au niveau agrégé par des changements dans la couverture par la presse de l'objet de vote.

Dans un registre assez proche on trouve l'interminable problème de l'effet de l'âge, par exemple, sur le vote pour l'EEE. L'âge en tant que tel reflète plusieurs facettes. Les analyses VOX montrent que les personnes âgées avaient une tendance plus forte à voter contre l'EEE (Figure 1). La question se pose immédiatement de savoir si c'est le vieillissement qui a poussé ces personnes à déposer un non fracassant dans l'urne, ou si c'est un effet de génération. Il se pourrait bien que les personnes ayant vécu la deuxième guerre mondiale et la période d'après-guerre, par leur vécu, ont une vue fortement différente sur une intégration supranationale. Ces

²⁵Le même genre de problème se pose aussi pour l'approche choisie par Buri et Schneider (1993). Voir aussi Markus (1988).

²⁶Notons que Kriesi (1994) tente une autre analyse portant sur l'utilisation de différentes sources d'informations. L'auteur essaie d'expliquer la décision de vote sur l'EEE en utilisant comme variables indépendantes une série de variables dichotomiques. Chacune mesure si la personne a eu recours à un type de média (radio, télévision, presse écrite, etc.) pour s'informer sur les positions et les enjeux d'une votation. Il n'est pas étonnant que les résultats soient relativement décevants: peu de médias semblent avoir un impact sur la décision de vote (Kriesi, 1994:60s). Dans la mesure où les journaux peuvent refléter des positions très différentes, l'impact des journaux peut aller dans un sens positif pour un individu et dans un sens inverse pour une autre. En total, les deux directions peuvent s'annuler et les coefficients estimés perdent toute leur signification substantielle. En contrôlant pour la région linguistique, Kriesi (1994:63) affine son analyse, mais ceci ne règle point les problèmes mentionnés.

deux effets bien distincts de l'âge, sont impossibles à décortiquer à l'aide d'une seule enquête. Dans notre exemple, qui porte sur un seul point dans le temps, génération et vieillissement sont en parfaite correspondance. Seuls des enquêtes longitudinales peuvent apporter des indices à des réponses possibles. Pour un vote bien particulier comme le vote sur l'EEE, ceci devient pratiquement une impossibilité.

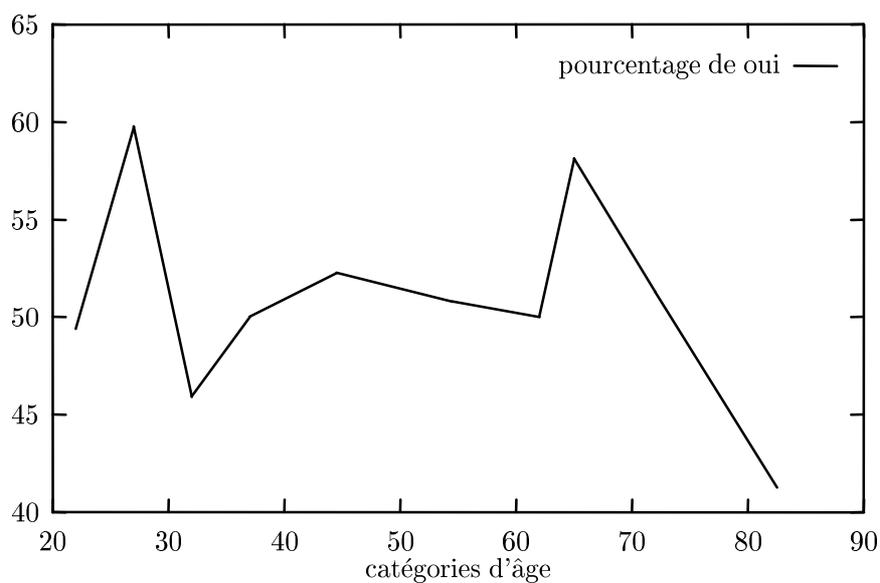


Figure 1: Pourcentage de oui et catégories d'âge

Un dernier exemple nous est fourni par Kramer (1983). En étudiant l'impact des politiques économiques du gouvernement sur l'électorat américain, cet auteur montre clairement que cette question de recherche nécessite le recours à des données agrégées. En règle générale, les politiques économiques d'un gouvernement ont un impact sur le bien-être économique des citoyens. Mais ce bien-être est aussi affecté par une multitude d'autres facteurs. Les augmentations de salaires, la perte du travail, un gain dans une loterie, tous peuvent avoir un impact considérable sur le bien-être économique. Mais ces facteurs sont largement individuels, et en partie dûs au hasard. En revanche, la politique économique d'un gouvernement est la même pour tout un pays et ne varie guère d'un individu à l'autre. L'impact d'une quelconque politique est en plus souvent le même pour des groupes d'individus. Donc, l'effet de la politique a de fortes chances d'être relativement faible face aux variations importantes au niveau individuel. Donc, l'estimation au niveau individuel pose problème.²⁷ Kramer montre même que les coefficients estimés à l'aide de

²⁷Markus (1988), utilisant une enquête par panel, montre toutefois que des possibilités existent

données individuelles sont entachés d'un biais pouvant être considérable. Un point similaire est relevé par Brown (1991: 43), en mettant surtout le doigt sur l'aspect changeant des comportements et attitudes des personnes:

“[t]he individual-level fallacy and the equilibrium fallacy . . . are, respectively, the danger of (1) extracting individuals from a context dependent process, and (2) extracting individuals or aggregates from a dynamic process. These fallacies generate problems at least as severe as the traditional bogeyman, the ecological fallacy.”

Ayant illustré avec plusieurs exemples les problèmes qui peuvent se manifester lors d'analyses au niveau individuel, il nous semble évident que ces derniers ne sont pas forcément la panacée. L'interprétation aveugle de la citation de Robinson (1950) est clairement erronée. Nous nous joignons donc à Huckfeldt et Sprague (1993: 286):

“As a logical matter, it is no more dangerous to infer individual behavior from aggregate level data than to infer any one person's behavior from individual data.”

3 Le niveau agrégé

La réaction à l'article de Robinson ne s'est guère fait attendre, mais son impact dans la littérature a été moindre. Peu de temps après la publication de l'article en question, Goodman (1953 et 1958) montrait sous quels postulats des regressions calculées au niveau agrégé pouvaient donner des indications précises sur les comportements des individus. Le postulat central en est que les relations au niveau individuel ne varient pas d'une unité d'agrégation à l'autre.²⁸ Ainsi, étudiant le transfert de voix du parti républicain au parti démocratique d'une votation à l'autre, la régression écologique est faisable, si dans les unités d'agrégation le même pourcentage d'anciens républicains votent démocrates, et vice-versa.

Pour des tableaux croisés d'autres méthodes ont été proposées et utilisées (Shively, 1969 Claggett et van Winden, 1993 et pour un survol Achen et Shively, 1994: ch. 1 et 8). Pour reprendre l'exemple des élections américaines, les résultats au niveau agrégé correspondent dans un certain sens à un tableau croisé. Une variable

de trouver des réponses à l'aide de données individuelles, combinées avec des indicateurs au niveau macro. En revanche, des problèmes d'estimation de paramètres se posent d'une manière importante. De nouveau, des solutions potentielles à ce problème sont fournies par des estimations de type “multi-level,” qu'on discutera ci-dessous.

²⁸Achen et Shively (1994: ch. 3,5 et 6) reprennent et discutent ces postulats tout en les élaborant.

est le vote lors de la première élection, l'autre le vote lors de la deuxième élection. Etant donné qu'on a disposition des données agrégées on connaît uniquement les pourcentages dans les marges et ignore les valeurs dans les cellules du tableau. Mais ces marges réduisent l'intervalle des valeurs possibles pour le pourcentage d'anciens républicains qui votent démocrates. Le principe de base étant de calculer les bornes inférieures et supérieures pour un paramètre au niveau individuel, sur la base des marges au niveau agrégé. En introduisant des hypothèses au niveau substantiel, ces intervalles, dans lesquelles le paramètre se situe par nécessité, peuvent être réduites considérablement (Shively, 1969 et Achen et Shively, 1995: ch. 8). Le recours à la programmation linéaire (Clagett et van Winden, 1993) peut aussi servir à réduire ces intervalles, parfois relativement étendus.

Toutefois, la réaction prédominante a été d'éviter l'inférence écologique en adoptant comme niveau d'analyse les unités d'agrégation. Lorsque le chercheur contrôle l'agrégation, ceci est certes une voie viable. Il peut regrouper les individus qui font partis d'un même groupe social. De cette sorte l'intérêt du chercheur passe du comportement individuel à celui de groupes sociaux. Dans le cas où ce contrôle sur l'agrégation échappe au chercheur, le changement de l'unité d'analyse doit rendre suspecte l'entreprise. Ainsi, pour situer la discussion dans le contexte des votations, il est clair que le vote est à la fois un phénomène individuel et collectif (Joye, 1987: 18s). Mais lorsque les électeurs pénètrent dans l'isoloir, ils se trouvent bien seuls face à leurs bulletins de vote. Nous voulons nullement affirmer que lorsque la porte se referme, que l'électrice ou l'électeur laisse tout son bagage et ses liens sociaux dehors. Mais ce bagage et ces liens méritent bien une définition soignée.

Malheureusement, cette définition fait souvent défaut. Par pur hasard, lorsque la chercheuse dispose de données communales, les acteurs pertinents pour sa recherche deviennent les communes: "Zurich a rejeté . . .". Son collègue qui travaille sur la même question, mais par malchance n'a que des données cantonales, affirme implicitement que les cantons sont les acteurs pertinents. Les deux oublient que ce n'est pas le fait d'avoir à disposition des données pour certaines unités, qui en font soudainement des acteurs sociaux. Tout acteur social est plus qu'une simple agrégation d'individus basée sur un critère géographique.²⁹ Gardant ceci à l'esprit montre bien qu'un grand nombre de recherches, qui se refusent de faire une inférence écologique, est vide de sens.

²⁹Voir à ce sujet la discussion de Achen et Shively (1995: ch. 1), de Sprague (1982) et d'Iversen (1991: 4s).

Dans le contexte suisse, il est indéniable que les cantons et les communes, par leur histoire et leurs institutions, sont des entités bien définies. L’attachement des Suisses à leur commune et, dans une moindre mesure, à leur canton est bien documenté (Melich, 1991: 33ss). Malgré ceci, il est peu convaincant de dire que Genève, Zurich ou n’importe quelle autre ville d’une certaine taille,³⁰ ait accepté un certain projet, comme si c’était un acteur. Donc, à notre avis cette échappatoire à l’inférence écologique n’est que peu crédible:

“When aggregate outcomes can reasonably be thought of as having been produced by “unitary actors,” it may be unproblematic to treat them as macrolevel data. In cases where aggregate outcomes arise from more complicated individual-level processes, as with votes for political parties, aggregate presidential popularity, or political protests, aggregate-level analysis of covariation are less obviously appropriate, although nevertheless quite common. This has led some authors to propose individual-level causal models, aggregate them, and develop techniques for estimating the underlying causal relationship. Both approaches have their place, but researchers are not always clear about which tack they are taking, and this can lead to confusion about whether they are testing individual-level causal theories or describing aggregate-level relationships.” (Bartels et Brady, 1993: 128f)

3.1 L’erreur écologique comme erreur de spécification

D’autres portes de sortie à l’inférence écologique n’ont pas tardé à s’ouvrir. L’article de Hanushek, Jackson et Kain (1974), qui a reposé la question de l’inférence écologique d’une manière plus fondamentale, y figure en prominence. Critiquant Robinson (1950), les trois auteurs mettent le doigt sur le problème fondamental de la spécification du modèle. Plus particulièrement ils avancent que le modèle que Robinson estime est fortement mal spécifié. Plus particulièrement, par le fait que des variables importantes sont omises de l’équation qu’il estime,³¹ des biais importants peuvent se produire. Puisque ces variables omises sont fortement corrélées avec la variable indépendante et la dépendante, le coefficient estimé est considérablement biaisée.

Par une respecification du modèle de Robinson (1950) en incluant le pourcentage

³⁰On peut y inclure aussi une grande partie des communes se situant dans une agglomération.

³¹Il faut noter que dans son article Robinson (1950) ne se sert que de coefficients de corrélation. Ceci revient toutefois à l’estimation d’une équation de régression avec une seule variable indépendante.

de jeunes inscrits à l'école primaire, Hanushek et al (1974) montrent que le problème de spécification est beaucoup plus important, que celui de l'inférence écologique. En effet, en contrôlant par cette variable importante, les résultats basés sur les deux niveaux d'analyses se rapprochent considérablement. Malgré le fait que les variables omises causent des biais à la fois au niveau individuel et agrégé, ils sont plus importants à ce dernier niveau. Ceci pour la simple raison que d'une manière générale l'agrégation élimine beaucoup de variations aléatoires des variables au niveau individuel. De cette manière les corrélations au niveau agrégé deviennent plus fortes. Un exemple frappant apparaît dans les corrélations rapportées par Vatter (1994: 27): La corrélation entre l'appartenance à la région linguistique (allemande) et le vote pour l'EEE est de 0.83 au niveau des communes, mais seulement de 0.27 au niveau individuel.

3.2 Les effets d'agrégation

Tout en mettant l'accent principal sur les problèmes de spécification du modèle au niveau individuel et au niveau agrégé, Hanushek et al (1974) ne mentionnent qu'en passant les effets de l'agrégation proprement dit. Plus particulièrement, l'agrégation d'observations à un niveau d'analyse plus élevé, peut poser deux problèmes bien distincts. Le premier a trait à l'estimation des erreurs standards des coefficients de régression. Celui-ci peut être réglé d'une manière relativement simple. Le deuxième, en revanche, peut poser plus de problèmes. En effet, lorsque l'agrégation échappe au contrôle du chercheur, il se peut que le modèle statistique au niveau agrégé ne corresponde pas à celui au niveau individuel.

Penchons-nous d'abord sur le premier problème qui ne concerne que les erreurs standards qui sont attachées aux coefficients de régression. Ici, il se montre que, sous l'hypothèse d'un modèle bien spécifié au niveau individuel avec deux variables indépendantes et des erreurs (U_i) distribuées selon la loi normale, on a au niveau individuel:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i \quad (1)$$

$$\text{var}(U_i) = \sigma^2 \quad \forall i \quad (2)$$

Si les données à disposition sont en revanche agrégées la situation change. Supposons qu'on a affaire à plusieurs unités spatiales, par exemple les communes. Ces dernières sont composées de nombre très variable de personnes. Si la commune k est

habité par N_k individus, et toujours sous l'hypothèse qu'aucun effet d'agrégation vient troubler la relation, on trouve que

$$\overline{Y}_k = \beta_0 + \beta_1 \overline{X}_{1k} + \beta_2 \overline{X}_{2k} + \overline{U}_k \quad (3)$$

$$\text{var}(\overline{U}_k) = \frac{\sigma^2}{N_k} \quad (4)$$

Aussi longtemps que le nombre de personnes dans chaque unité varie d'une unité à l'autre, la variance des erreurs n'est point constante à travers les différentes observations:

$$\text{var}(\overline{U}_i) \neq \text{var}(\overline{U}_j) \quad \forall i, j \quad i \neq j \quad \text{et} \quad N_i \neq N_j \quad (5)$$

La violation du postulat de la variance constante des erreurs n'a pas d'impact sur la valeur attendue des coefficients estimés. En revanche, les erreurs standards estimés pour ces derniers sont biaisées. Etant donné que les auteurs travaillant avec les données agrégées mettent beaucoup d'accent sur les tests de significativité, une correction s'impose: la régression pondérée, qui fait partie de la méthode généralisée des moindres carrés, offre la solution.³²

Reprenant l'exemple de la partie précédente et utilisant la correction discutée au préalable on peut montrer que le problème de spécification est toute aussi crucial au niveau agrégé. De nouveau un modèle utilisant comme seule variable indépendante le fait d'habiter en ville montre un effet important d'être citoyen.³³ Si l'on contrôle cette relation par les autres variables, cet effet disparaît quasi totalement.

Tandis que le problème de la variance non-constante des termes d'erreurs peut être corrigé relativement facilement, il en est tout autre en ce qui concerne le deuxième effet de l'agrégation. Ce problème vient du fait qu'il n'est pas suffisant

³²A quelques exceptions près (Weck-Hannemann, 1990 et Nabholz et Vatter, 1993) ces corrections ne sont guère faites ou alors point rapportées. L'utilisation des coefficients standardisés ou des procédures d'estimation de type "stepwise" sont dans ce cas-là encore plus problématique que dans les situations normales (King, 1986). De même l'élimination des petites communes (par exemple Vatter, 1994) n'est guère souhaitable et ne règle nullement le problème de la variance non-constante des erreurs.

³³La variable dépendante a été transformée par l'opération suivante: $\ln \frac{P}{1-P}$, où P est la proportion des personnes ayant accepté l'EEE. Par cette opération la variable dépendante peut varier entre $-\infty$ et $+\infty$. En plus on a procédé à une régression pondérée. Les poids retenus sont égales à $N \frac{P}{1-P}$ (Hanushek et Jackson, 1977: 193), où N correspond aux nombres de votants dans une commune donnée, et P à la proportion de personnes ayant voté pour l'EEE. Gauche correspond à la proportion de votes obtenue par le parti socialiste lors des élections fédérales de 1991, ville est égale à un pour les 33 agglomération défini par l'Office fédérale de la statistique, romande est égale à un pour tous les communes où le pourcentage de personnes de langue maternelle française dépasse les 50 pourcents, et éducation élevée correspond à la proportion ayant fréquenté l'université, une école préparant à l'obtention de la maturité ou une école professionnelle supérieure.

| | b | erreur standard | b | erreur standard | b | erreur standard |
|------------------|-------|-----------------|---------|-----------------|--------|-----------------|
| ville | .528 | .033 | .106 | .029 | .428 | .019 |
| éducation élevée | | | 19.1951 | 0.600 | 6.508 | .421 |
| gauche | | | 2.578 | .136 | 1.541 | .086 |
| romande | | | | | 1.401 | .021 |
| (Constant) | -.295 | .019 | -1.588 | .037 | -1.190 | .024 |
| Standard Error | | 4.279 | | 4.279 | | 2.149 |
| n | | 2890 | | 2890 | | 2890 |

Table 7: Trois modèles d'explication du vote sur l'EEE (régression écologique)

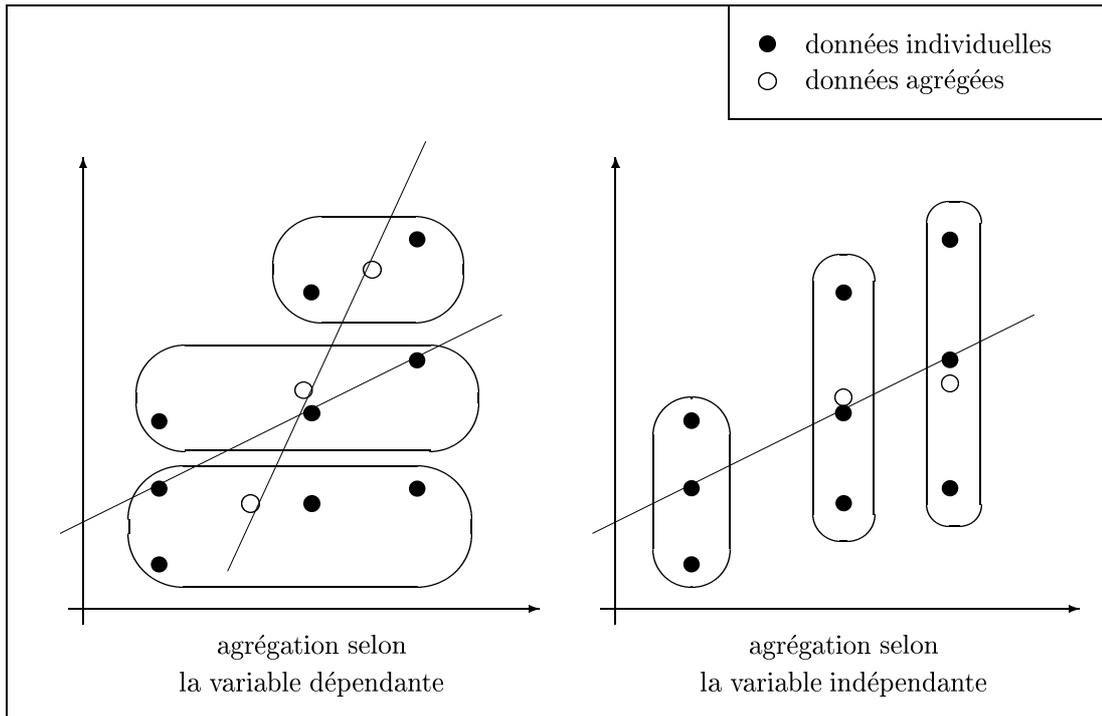


Figure 2: Deux types d'agrégation

d'agréger un modèle bien spécifié au niveau individuel, pour aboutir à des coefficients estimés correctement au niveau agrégé. Ceci dépend fortement de la manière selon laquelle les unités d'observation (les individus) ont été regroupées. Si ce regroupement est lié d'une certaine manière à la variable dépendante, des biais dus à l'agrégation sont inévitables:³⁴

“ . . . [W]hen individual level models are properly specified, the consequences of aggregation on cross-level inferences depend on the grouping process itself. Two forms of grouping - grouping by the independent variable and random grouping - do not affect the specification of macrorelations. Parameter estimates at both levels of analysis include

³⁴Voir aussi Achen et Shively, 1995: ch.1.

no specification bias. But two other forms - grouping by the dependent variable and grouping by a variable related to both the independent and dependent variable - do generate specification bias in macrorelations.” (Langbein et Lichtman, 1978: 21)

Nous pouvons illustrer ces effets d’agrégation dans la figure 2, où nous reproduisons une simple relation bivariée. A gauche, lorsque les neuf observations sont agrégées en fonction de la variable dépendante, l’estimation de la pente au niveau agrégé est fortement biaisée. En revanche, sur la droite de graphique, il se montre qu’une agrégation basée sur une variable explicative n’induit pas de biais.

Nous pouvons tenter d’illustrer ce même problème avec un exemple construit sur la base des données VOX. Au lieu d’avoir à disposition les données sur le vote sur l’EEE au niveau individuel, nous agrégeons ces données au niveau cantonal. Il est fort probable que cette agrégation est liée d’une manière ou d’une autre avec la variable dépendante. En témoignent les différences marquantes des pourcentages d’acceptation entre les différents cantons. En estimant un modèle explicatif très simple nous arrivons aux résultats qui figurent dans la colonne au milieu du tableau 8. Ces estimations sont basées sur 24 unités d’observations, étant donné que dans deux cantons aucune personne n’a été interrogée pour le sondage VOX.³⁵

Dans un deuxième temps nous avons de nouveau créé 24 unités d’observations sur la base des personnes interrogées pour le sondage VOX. Mais cette fois-ci l’agrégation a été faite d’une manière aléatoire. Plus précisément, l’ensemble des personnes interrogées ont été réparties aléatoirement dans 24 groupes. Etant donné que le résultat basé sur une telle agrégation dépend fortement du générateur de nombres aléatoires, nous avons répété l’opération dix fois. Par conséquent, les coefficients qui apparaissent dans la dernière colonne du tableau 8 sont des moyennes calculées sur les dix estimations. Ils correspondent ainsi dans un certain sens aux valeurs attendues d’une régression basée sur une agrégation aléatoire, donc une qui n’est pas liée à la variable dépendante.

Toujours sous l’hypothèse d’un modèle bien spécifié, les résultats dans la dernière colonne du tableau 8 devraient correspondre aux effets au niveau individuel (première colonne du tableau). En revanche, ceux de la deuxième colonne du tableau, vu le fait que l’agrégation a été faite d’une manière qui est potentiellement et presque

³⁵Notons que nous avons estimé cette équation en pondérant en fonction du nombre de personnes interrogées dans chaque canton et la proportion de votes en faveur de l’EEE. Pour plus de détails voir ci-dessus.

| | individuel | | agrégation par canton | | agrégation aléatoire (moyenne sur 10 estimations) | |
|-------------------|------------|-----------------|-----------------------|-----------------|--|-----------------|
| | b | erreur standard | b | erreur standard | b | erreur standard |
| ville | 0.279 | 0.220 | 0.120 | 0.422 | 0.466 | 0.954 |
| éducation élevée | 1.005 | 0.242 | 3.335 | 1.622 | 0.686 | 0.977 |
| romande | 0.821 | 0.257 | 1.704 | 0.333 | 1.016 | 0.978 |
| gauche | 0.946 | 0.233 | 1.208 | 0.805 | 0.523 | 0.714 |
| Constant | -0.646 | 0.159 | -1.475 | 0.412 | -0.195 | 0.605 |
| erreur standard | | | | 0.950 | | 0.730 |
| -2 Log-likelihood | | 550.829 | | | | |
| n | | 434 | | 24 | | 24 |

Table 8: Les effets d'agrégation

certainement liée avec la variable dépendante, devraient être biaisé. Effectivement, nous observons des changements importants dans les coefficients estimés. Avec le jeux de données constitué à l'aide d'une agrégation aléatoire il se montre que tous les effets sont assez proche des coefficients estimés au niveau individuel.³⁶ En revanche, lorsque nous nous servons du jeux de données basé sur l'agrégation en fonction du pourcentage de oui, les résultats changent considérablement. Mais ces effets sont tous dûs à une agrégation qui est liée à la variable dépendante. Seulement en ajoutant des variables contrôlant pour cet effet de l'agrégation les estimations deviennent non-biaisées:³⁷ Ces différences deviennent encore plus visible dans le graphique suivant.

Du graphique (3) ressort clairement que les estimations basées sur une agrégation aléatoire sont plus proche de celle basée sur les données individuelles. En revanche, à la fois pour l'agrégation au niveau cantonal et l'estimation sur la base de données communales, des biais importants s'introduisent. Plus particulièrement le coefficient qui indique le changement dans le vote en fonction d'une éducation élevée, est fortement affecté. Les résultats basés sur l'estimation de données agrégées au niveau cantonal ou communales 7 montrent un biais important. Dans les deux cas, l'impact d'un niveau d'éducation élevée est fortement surestimé. Comparant uniquement l'agrégation aléatoire à celle au niveau cantonal (Tableau 8) montre que le coefficient dans le deuxième cas est trois fois plus important que dans le premier.

“Intuitively speaking, . . . aggregate cross-sectional regression is unbiased only when the micro-level specification is good enough to make geographic location useless as a predictor of the dependent variable.”

(Achen and Shively, 1995: ch.4)

³⁶La différence entre les deux colonnes ne dépasse jamais la valeur d'une erreur standard.

³⁷Ceci est partiellement une justification pour les méthodes présentées ci-dessous.

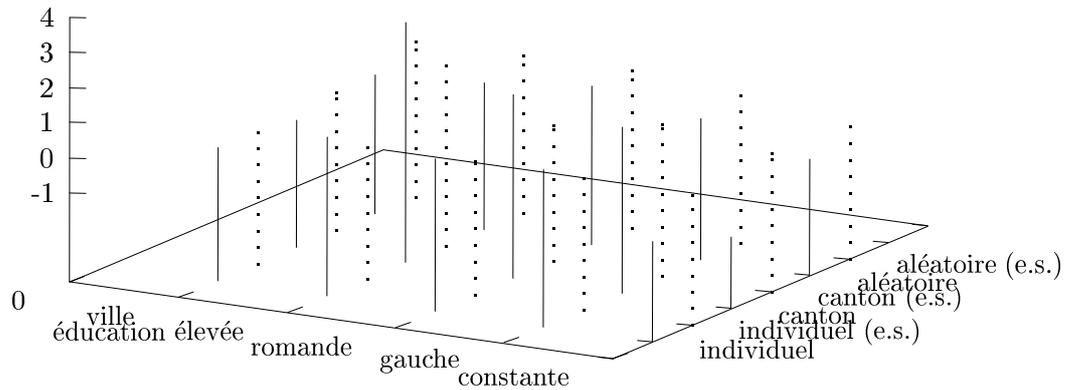


Figure 3: Comparaison de deux types d'agrégation

4 Une analyse “multilevel”

Dans les trois sections précédentes, nous avons tenté de montrer que les données individuelles ne constituent pas la seule base possible de l'explication du comportement individuel. Les données agrégées sont parfois utiles, voire nécessaires. Dans cette dernière section, nous allons présenter une approche qui permet de combiner les deux niveaux dans l'analyse. Les techniques “multilevel” permettent de mesurer l'influence du contexte sur les relations entre variables individuelles.

4.1 Qu'est-ce qu'une analyse multilevel?

La plupart des chercheurs en science sociale s'accordent à reconnaître l'existence de diverses hiérarchies à plusieurs niveaux au sein de la société: les électeurs, au niveau 1, votent dans une circonscription, au niveau 2, qui, elle-même appartient au niveau 3 à une région. De même, les résidents suisses qui doivent remplir le questionnaire du recensement appartiennent d'abord à un ménage, qui se trouve dans un secteur de recensement au sein d'une commune d'un canton particulier. Il est à noter que ces hiérarchies ne sont pas forcément géographiques; elles peuvent par exemple se manifester dans le cas d'une enquête panel où une réponse à un moment donné au

niveau 1 appartient à un individu au niveau 2.

Dans le cas du vote, de nombreuses recherches ont montré l'influence du contexte sur le comportement individuel.³⁸ Dans ce texte, l'influence de la région linguistique en est une illustration. De plus, comme nous l'avons vu, une analyse qui utilise uniquement des données agrégées risque l'erreur écologique ou le biais d'agrégation (Roberts and Burnstein, 1980) tandis qu'une étude essentiellement individuelle peut s'exposer à au contraire à l'erreur atomistique (atomistic fallacy: Alker, 1969). L'intérêt d'une approche où il est possible de combiner les données agrégées et individuelles est donc évident.

La réflexion théorique sur l'influence des hiérarchies dans l'analyse sociale existe déjà depuis plus d'une vingtaine d'années (Jones and Duncan, 1994: 2). L'analyse multilevel, en tant que technique statistique servant à estimer l'importance relative des facteurs contextuels et individuels n'a cependant été développée dans la seconde moitié des années 80. C'est dans le domaine de la sociologie de l'éducation que sont apparues les techniques statistiques qui permettent d'effectuer des analyses multi-niveaux.³⁹

4.2 Un modèle multilevel simple

Afin de comprendre les principes d'une analyse multilevel, nous allons ici développer un modèle multi-niveaux avec des données imaginaires.⁴⁰ Pour simplifier cette première explication, nous nous contenterons d'imaginer ici un modèle simple qui explique la probabilité de vote pour l'EEE à partir de l'orientation idéologique sur une échelle droite-gauche.⁴¹ Deux niveaux seront considérés: le niveau individuel et celui des cantons.

En nous limitant au niveau individuel, on pourrait estimer ce modèle à l'aide d'une régression linéaire simple représentée par l'équation 6; y_{ij} représente la probabilité de voter oui de l'individu i du canton j et x_{ij} la position sur l'échelle droite-gauche de cet individu en tant qu'écart à la moyenne,⁴² il s'agit des tests de signifi-

³⁸Huckfeldt et Sprague (1993) présentent un survol de cette littérature. Caramani (1994) présente une application plus récente.

³⁹Plusieurs logiciels sont aujourd'hui disponibles, chacun présente des caractéristiques distinctes. Ici, nous avons choisi d'utiliser ML3 qui repose sur une méthode d'estimation développée par l'un des pionniers du domaine, Goldstein (1986, 1987) et basée sur la généralisation des moindres carrés.

⁴⁰Cette façon de faire est assez commune dans les textes qui utilisent une analyse multilevel, voir par exemple Jones, Johnston and Pattie (1992: 348ss) ou Woodhouse (1993).

⁴¹Un tel modèle imaginaire présente deux avantages principaux: La variable dépendante est une probabilité et peut donc être considérée comme continue, et on peut faire le même postulat pour la variable indépendante.

⁴²Centrer sur la moyenne est utile pour deux raisons. D'abord l'interprétation de la constante est grandement facilitée. La seconde raison est plus spécifique à l'analyse multilevel: elle tient à

cation le centre de l'échelle. a mesure la probabilité attendue pour les personnes qui se situent au centre de l'échelle et b l'augmentation attendue de cette probabilité pour chaque unité d'écart par rapport au centre de l'échelle. e_{ij} , le terme d'erreur, donne une indication de l'écart entre la probabilité prédite réelle.

$$y_{ij} = a + bx_{ij} + e_{ij} \quad (6)$$

Dans la tradition des analyses multilevel, nous distinguerons la partie fixe de la partie aléatoire de cette équation. La première ne varie pas d'un individu à l'autre: il s'agit des coefficients a et b qui sont communs à tous les individus. Ces deux coefficients définissent la droite de régression telle qu'elle apparaît dans la figure 4-a. On y observe que plus l'individu se situe à gauche, plus sa probabilité de voter oui est grande. La pente de la droite est mesurée par b et le lieu d'intersection entre la verticale qui part de "centre" et la droite de régression a une valeur de a . La partie aléatoire est pour l'instant uniquement composée du terme d'erreur e_{ij} qui est propre à chaque observation. Une observation donnée pourrait être représentée par un "point" à l'intersection des deux valeurs de cet individu sur les axes. Le terme d'erreur mesure la distance verticale entre ce point et la droite de régression.

Selon les postulats de la régression linéaire, le terme d'erreur est distribué selon la loi normale avec une moyenne de 0 et une variance σ_e^2 constante. Dans un tel postulat, nous faisons d'abord l'hypothèse que la relation entre l'orientation idéologique et le vote est linéaire, c'est-à-dire, identique à tous les niveaux de la variable indépendante. Ensuite, ce postulat de la variance constante implique l'hypothèse que la relation est la même à l'intérieur de chaque canton. C'est évidemment ce deuxième postulat qui va être remis en cause par le modèle multilevel.

Notre modèle est, pour l'instant défini au seul niveau individuel. Comment introduire les cantons? Dans un premier temps, nous nous contenterons de postuler que la probabilité de voter oui est différente dans chaque canton, mais que la pente de la relation y est uniforme. Ce modèle, dit de "random intercepts" (Jones, Johnston and Pattie, 1992: 350) est présenté dans l'équation 7.

$$y_{ij} = a_j + bx_{ij} + e_{ij} \quad (7)$$

La différence avec la première équation concerne le coefficient a_j qui varie maintenant d'un canton à l'autre. Cela signifie que la probabilité prédite pour les indi-
la significativité de la variance au deuxième niveau (cf. noteb).

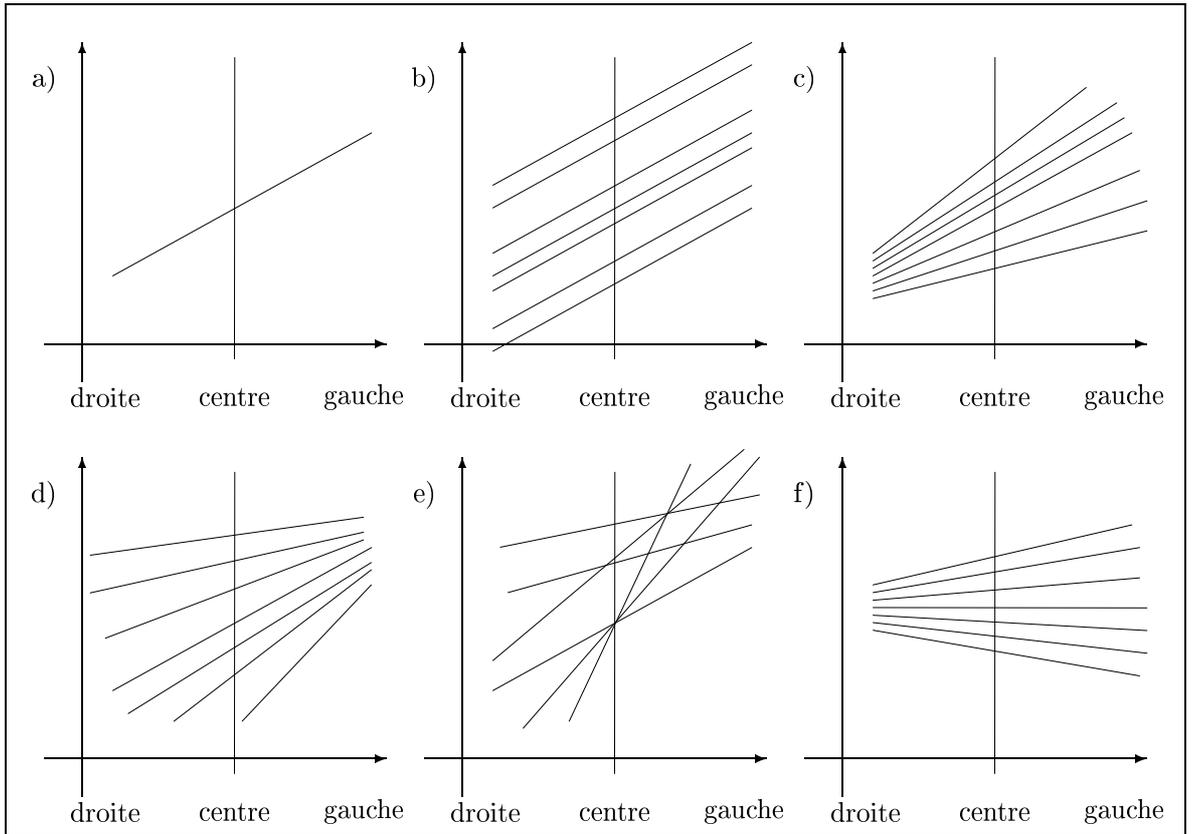


Figure 4: Variantes de relations entre position idéologique et probabilité de vote pour l'EEE

vidus positionnés au centre peut varier d'un canton à l'autre. En nous référant à la distinction entre partie fixe et aléatoire, nous décomposons dans l'équation 8 le coefficient a_j entre a qui renvoie à l'équation 6 et u_j qui représente, pour le canton j , l'écart entre a et a_j .

$$y_{ij} = a + bx_{ij} + (u_j + e_{ij}) \quad (8)$$

Dans cette équation, nous avons donc deux termes dans la partie aléatoire: u_j et e_{ij} . u_j peut en fait être décrit comme le terme d'erreur aléatoire du niveau 2, à savoir des cantons. Nous faisons ici le même postulat quant à la variance de ce terme qu'au premier niveau; elle peut donc être résumée parfaitement par σ_μ^2 .

La figure 4-b traduit graphiquement ce modèle: chaque canton a une droite de régression de pente identique, mais de niveau différent. Nous postulons donc que la relation est la même à l'intérieur de chaque canton. Que ce passe-t'il si nous supprimons ce postulat?

Les figures 4-c, 4-d et 4-e illustrent des modèles basés sur l'élimination de cette contrainte. Nous nous contenterons à présent de dire que les niveaux et les pentes cantonales sont différentes dans ces graphiques.

Du point de vue mathématique, la procédure est similaire à celle de l'équation 8. Pour permettre la variation de la pente, il faut créer un terme b_j qui varie selon le canton. Celui-ci peut être décomposé en b qui représente la pente moyenne et v_j , l'écart entre la pente moyenne et celle du canton j . L'équation 9 est issue de ce nouveau modèle.

$$y_{ij} = a + bx_{ij} + (v_j x_{ij} + u_j + e_{ij}) \quad (9)$$

Là encore, nous avons ajouté un nouveau coefficient, v_j , dans la partie aléatoire. Dans la terminologie multilevel, ce modèle est appelé "fully random" ou "random slopes two-level model" (Jones, Johnston and Pattie, 1992: 351). Selon le même postulat que précédemment, la variance de v_j , σ_v^2 , permet de résumer parfaitement ce terme.

La différence entre les trois graphiques 4-c, 4-d et 4-e résulte de la covariance estimée entre u_j et v_j . Si cette covariance est positive, cela signifie que plus le niveau cantonal u_j est élevé, plus la pente cantonale v_j est forte. Ce premier cas est illustré dans la figure 4-c. Le cas d'une covariance négative est montré dans la figure suivante où la pente diminue lorsque le niveau de la droite augmente. En dernier lieu, il convient de distinguer la situation d'une covariance nulle telle qu'elle apparaît dans la figure 4-e où les pentes cantonales sont aléatoirement différentes.

Ainsi, dans un modèle multi-niveaux totalement aléatoire, il faut estimer trois paramètres de plus que dans une régression simple: σ_u^2 , σ_v^2 et σ_{uv} . En général, on procède en deux temps: d'abord en effectuant un modèle à ordonnées à l'origine variables, ensuite en spécifiant un modèle totalement aléatoire. Si la variance σ_u^2 du premier modèle n'est pas significative, alors les cantons n'ont pas d'influence sur la probabilité attendue d'un vote positif. Dans le cas contraire, il convient d'estimer le second modèle qui nous dira si la pente de la relation varie significativement⁴³ d'un canton à l'autre. L'interprétation des coefficients fixes est identique à celle d'une régression normale, de plus, il est facile de calculer les valeurs prédites tant au niveau des individus que des cantons (Woodhouse, Rasbash, Goldstein and Yang, 1993: 39-40).

⁴³C'est ici qu'apparaît l'utilité de centrer l'orientation: la significativité est mesurée par rapport à une valeur qui fait sens.

4.2.1 Quels sont les avantages du multilevel?

On pourrait objecter à ce stade que ce que l'analyse multilevel nous apporte, à savoir une estimation de la variation cantonale, pourrait tout aussi bien se calculer à l'aide d'une régression simple où l'on introduirait des variables dichotomiques représentant les cantons.⁴⁴

Sans entrer dans les détails, on peut répondre que dans le cas d'un modèle simple à random intercepts, il faudrait créer 25 variables dichotomiques et estimer leurs coefficients si on voulait utiliser la régression simple. Dans le cas du multilevel, un seul coefficient, σ_u^2 , doit être estimé. En plus, l'estimation des valeurs attendues sera meilleure dans ce deuxième cas pour les cantons où le nombre d'observations est faible parce que la technique d'estimation utilisée dans le multilevel tend à ramener le coefficient de ces cantons à la moyenne suisse. On utilise donc l'information des cantons où l'on dispose de nombreuses observations pour corriger les estimations des autres (Woodhouse, Rasbash, Goldstein and Yang, 1993: 42).

On peut dire en fait que l'analyse multilevel se base sur la distribution statistique des valeurs cantonales pour estimer les pourcentages par canton alors que la régression normale utilise uniquement les données individuelles. En simplifiant quelque peu, il est possible de comparer la technique d'estimation du multilevel à deux régressions successives, la première sur les données individuelles et la seconde sur les valeurs agrégées au niveau cantonal. C'est la raison pour laquelle il faut disposer d'un nombre suffisant d'unités au deuxième niveau.⁴⁵

Dans le cas d'un modèle à random slopes, la régression nécessite l'estimation de 25 coefficients supplémentaires pour les interactions. Ceci signifie que la régression normale nécessite 47 degrés de libertés de plus que le multilevel qui n'en a besoin que de trois. La précision des estimations cantonales est aussi meilleure, pour les mêmes raisons que dans le cas précédent.

4.2.2 Analyse multilevel d'un modèle simple du vote sur l'EEE

Pour voir ce que l'analyse multilevel peut nous apporter, nous effectuons dans cette dernière partie l'analyse multilevel d'un modèle du vote sur l'EEE. Dans notre problématique, la variable dépendante est dichotomique, il s'agit donc d'effectuer une régression logistique multilevel. Cette procédure est plus complexe (Paterson

⁴⁴Pour une discussion approfondie des avantages du multilevel par rapport à la régression simple, v. Aitkin et Longford (1986).

⁴⁵Paterson et Goldstein (1993) conseillent d'utiliser au moins 25 observations au niveau 1 dans 25 unités au niveau 2.

1991,1993i) que l'analyse qui est décrite dans la première partie de cette section. Au vu des résultats que nous avons obtenus, nous pourrions cependant nous contenter d'une interprétation sommaire. Le tableau 9 présente les résultats de l'estimation multilevel de deux modèles simples du vote sur l'EEE. Les coefficients de la partie fixe du modèle ont la même signification que dans une régression logistique normale, il s'agit du logarithme du rapport entre la probabilité de voter oui et celle de voter non. La partie aléatoire comprend uniquement les estimations de la variance des termes d'erreur aux deux niveaux considérés.

| | A | | B | |
|--|--------|-----------------|--------|-----------------|
| | b | erreur standard | b | erreur standard |
| FIXE | | | | |
| ville (ml) | 0.237 | 0.152 | 0.210 | 0.152 |
| (ols) | 0.299 | 0.155 | 0.338 | 0.149 |
| romande (ml) | 1.198 | 0.223 | | |
| (ols) | 1.549 | 0.200 | | |
| éducation élevée (ml) | 0.738 | 0.163 | 0.676 | 0.155 |
| (ols) | 0.836 | 0.179 | 0.691 | 0.173 |
| gauche (échelle) (ml) | 0.758 | 0.153 | 0.717 | 0.145 |
| (ols) | 0.807 | 0.167 | 0.808 | 0.161 |
| constante (ml) | -1.043 | 0.151 | -0.743 | 0.173 |
| (ols) | -0.864 | 0.136 | -0.557 | 0.126 |
| ALEATOIRE | | | | |
| Niveau 2 (Cantons) $\sigma_{u_j}^2$ | 0.072 | 0.055 | 0.319 | 0.134 |
| Niveau 1 (Individus) $\sigma_{e_{ij}}^2$ | 0.992 | 0.045 | 0.939 | 0.043 |
| n | | 862 | | 862 |
| -2 Log-Likelihood | | 1066.9 | | 1136.1 |

Table 9: Deux modèles multilevel du vote sur l'EEE

Dans le premier modèle, nous avons introduit le fait d'habiter en ville, l'appartenance à la Suisse romande, la possession d'un diplôme d'études supérieures ainsi que le positionnement à gauche sur l'échelle gauche-droite comme autant de variables explicatives dichotomiques. Cette spécification correspond à un modèle random intercepts tel qu'il a été décrit plus haut: les relations entre les variables indépendantes et la dépendante sont supposées identiques au sein de chaque canton. Seule la constante, c'est-à-dire la valeur attendue de la catégorie de référence, peut varier d'un canton à l'autre.

Dans la partie fixe du modèle, les coefficients multilevel diffèrent assez peu de ceux d'une régression logistique ordinaire. La seule différence notable concerne l'appartenance à la romandie. Il est normal que le coefficient du multilevel soit plus bas parce que la variance entre les cantons explique une partie de l'influence de ce facteur.

L'information la plus intéressante de ce premier modèle se trouve néanmoins au niveau de la variance estimée au niveau des cantons. Il semble en effet que les variations inter-cantoniales ne soient pas significatives, et que le recours à une

analyse multilevel soit donc superflu.

Un tel constat s'explique aisément à l'aide du deuxième modèle dans lequel la variable "romande" ne figure pas. Sans nous attarder sur les autres paramètres, il apparaît clairement dans ce deuxième modèle que la variation entre les cantons est significative. Qu'est-ce que cela signifie? Tout simplement que les différences entre les cantons s'expliquent presque entièrement par la région linguistique.

Ainsi, nous avons quand même tiré une information intéressante de résultats qui, à priori, semblaient fort décevants. L'influence du contexte ne s'exerce donc sur le vote pour l'EEE qu'au sein des régions linguistiques. Au niveau cantonal, il semble que l'essentiel des différences s'explique par les caractéristiques individuelles des citoyens.

5 Conclusion

"Toutefois, le but de ce papier sera atteint s'il empêche dans le futur des discussions vides de sens où des positions dogmatiques sur l'inférence écologique s'affrontent. De même nous espérons qu'il stimule pour des recherches similaires l'utilisation soignée des méthodes appropriées et des échanges et collaborations entre les tenants des deux approches."
(Brunner et Hug, 1994)

En concluant avec une paraphrase de la célèbre citation de Robinson (1950: 357), nous voulons réitérer notre souci de pousser plus loin le débat autour de l'erreur écologique. Comme nous l'avons montré, cette erreur, tout en étant importante, n'est pas la seule erreur qui peut se produire dans des analyses quantitatives en science politique ou en sciences sociales en général. D'autres niveaux d'analyse connaissent aussi des "erreurs" qui ne sont pas forcément moins problématiques. Donc, un respect mutuel entre les chercheurs utilisant des données individuelles et les spécialistes en données agrégées est nécessaire. Toutefois, ceci ne signifie nullement qu'en choisissant une approche on soit dispensé de prendre en considération les problèmes, parfois inévitables, inhérentes à celle-ci.

Tandis que les problèmes des données individuelles sont relativement bien connus, ils ont souvent été minimisés face à l'erreur écologique. Cette dernière a été diabolisée, bien que ses causes soient connues et discutées dans la littérature depuis longtemps, elles n'ont peu été prises en considération dans les études empiriques. Toutefois, on connaît bien les raisons pour lesquelles une relation au niveau agrégé

peut différer d'une autre qui se base sur le niveau individuel. D'abord, les biais causés par des erreurs de spécification sont amplifiés au niveau agrégé. Ensuite, et plus important, les biais sont fortement liés au type d'agrégation. Lorsque l'agrégation n'est pas liée avec le phénomène pour lequel on essaie de trouver une explication, des biais sont inexistantes. Dans tous les autres cas, des problèmes se posent. Vu le fait que l'agrégation est rarement sous le contrôle du chercheur, il est évident que ce risque est omniprésent.

Dans ce contexte, l'analyse multi-level, en mettant fortement l'accent sur les différences entre les unités agrégées, fournit un instrument important pour combiner les analyses au niveau individuel et agrégé. Pour les analyses des votations en Suisse cette méthode pourrait servir comme un outil important pour pousser plus loin les analyses faites à la fois sur la base des enquêtes VOX et les données communales et cantonales. Avec un exemple, pourtant simple, nous avons tenté d'illustrer l'utilité de cette approche et espérons ouvrir la porte pour des tentatives ultérieures, certes plus développées et sophistiquées.

6 Bibliographie

- Achen, Christopher H.; Shively, W. Philips 1995 (à paraître). *Cross-Level Inference*. Chicago: University of Chicago Press.
- Aitkin, M.; Longford, N. 1986. Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society, Series A* 149: 1-43.
- Aldrich, J.H.; Nelson, F.D. 1984. *Linear Probability, Logit, and Probit Models*. Beverly Hills : Sage Publications.
- Alker, H.S. 1969. A Typology of Ecological Fallacies. in Dogan, M. and Rokkan, S. (eds) *Quantitative Ecological Analysis*. Cambridge: MIT Press.
- Bartels, Larry M.; Brady, Henry E. 1993. The State of the Quantitative Political Methodology. in Finifter, Ada W. (ed.) *Political Science: The State of the Discipline*. Washington: American Political Science Association.
- Berglund, Sten; Thomsen, Soren Risbjerg (eds.) 1990. *Modern Political Ecological Analysis*. Abo: Abo Akademis Förlag.
- Blau, Peter M. 1993. Multilevel Structural Analysis. *Social Networks*. 15(2) 201-215.
- Brunner, Matthias; Hug, Simon 1994. *EEE et EEE. Essai sur l'erreur écologique et le vote sur l'Espace économique européen*. Balstahl: Papier présenté au congrès de l'Association Suisse de Science Politique.
- Buri, Christof; Schneider, Gerald 1993. Gründe und Scheingründe für das Schweizer Abstimmungsverhalten. *Revue suisse de sociologie*. 19. 389-417.
- Claggett, William; Wingen, John Van 1993. An Application of Linear Programming to Ecological Inference: An Extension of an Old Procedure. *American Journal of Political Science*. 37(2) 633ff.
- Converse, Jean M.; Presser, Stanley 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills: Sage Publications.
- Erbring, Lutz 1990. Individuals Writ Large: An Epilogue on the "Ecological Fallacy" *Political Analysis*. 1. 235-269.
- Erbring, Lutz; Young, Alice A. 1979. Individuals and Social Structure: Contextual Effects as Endogenous Feedback. *Sociological Methods and Research*. 7. 396-430.
- Feldman, S. 1991. What Do Survey Questions Really Measure? *The Political Methodologist*. 4. 8-12.

- Goertz, Gary D. 1994 (à paraître). *Contexts of International Politics*. New York: Cambridge University Press.
- Goldstein, Harvey 1991. Nonlinear Multilevel Models. *Biometrika*. 78(1) 45-51.
- Goldstein, Harvey; Rasbash, Jon 1992. Efficient Computational Procedures for the Estimation of Parameters in Multilevel Models Based on Iterative Generalised Least Squares. *Computational Statistics & Data Analysis*. 13(1) 63-71.
- Goldstein, H. 1986. Multilevel Mixed Linear Model Analysis Using Iterative Generalised Least Squares. *Biometrika*. 73: 43-55.
- Goldstein, Harvey 1987. *Multilevel Models in Educational and Social Research*. London: Charles Griffin.
- Goodman, Leo 1953. Ecological Regression and the Behavior of Individuals. *American Sociological Review*. 18. 663-664.
- Goodman, Leo 1957. Some Alternatives to Ecological Correlation. *American Journal of Sociology*. 64. 610-625.
- Gruner, Erich 1987. Die direkte Demokratie in der Bewährungsprobe. *Annuaire suisse de science politique*. 27. 283-314.
- Gruner, Erich; Hertig, Hans-Peter 1983. *Der Stimmbürger und die "neue" Politik*. Bern: Haupt.
- Hanushek, Eric A.; Jackson, John E. 1977. *Statistical Methods for Social Scientists*. New York: Academic Press.
- Hanushek, Eric A.; Jackson, John E.; Kain, John F. 1974. Model Specification, Use of Aggregate Data, and the Ecological Correlation Fallacy. *Political Methodology*. 1. 89-107.
- Huckfeldt, Robert; Sprague, John 1993. Citizens, Contexts, and Politics. in Finifter, Ada W. (ed.) *Political Science: The State of the Discipline*. Washington: American Political Science Association.
- Hug, Daniel; Kummer, Lorenz 1994. *Democratic Legitimation vs. International Reliability and Internal Cohesion. The Swiss Experience with Foreign Policy Referendums*. Madrid: Paper prepared for presentation at the Joint Session of Workshops of the ECPR.
- Hug, Daniel; Kummer, Lorenz; Vatter, Adrian 1993. *Die Volksabstimmung über den Beitritt der Schweiz zum Europäischen Wirtschaftsraum (EWR) vom 6. Dezember 1992. Eine Analyse der Gemeinderesultate, Zusammenfassung*. Bern: Bundesamt für Statistik.
- Hug, Daniel; Kummer, Lorenz; Vatter, Adrian 1993. *Die Volksabstimmung über den*

- Beitritt der Schweiz zum Europäischen Wirtschaftsraum (EWR) vom 6. Dezember 1992. Eine Analyse der Gemeinderesultate.* Bern: Forschungszentrum für schweizerische Politik, Universität Bern.
- Iversen, Gudmund R. 1991. *Contextual Analysis.* Newbury Park: Sage Publications.
- Jones, Kelvyn 1993. “*Everywhere is Nowhere*”: *Multilevel Perspectives on the Importance of Place.* Portsmouth: University of Portsmouth.
- Jones, Kelvyn; Duncan, Craig 1994. *People and Places: The Multilevel Model as a General Framework for the Quantitative Analysis of Geographical Data.* Southampton: Center for Survey Data Analysis, University of Southampton.
- Jones, Kelvyn; Johnston, John; Pattie, C.J. 1992. People, Places and Regions: Exploring the Use of Multi-Level Modelling in the Analysis of Electoral Data. *British Journal of Political Science.* 22. 343-380.
- Joye, Dominique 1984. *Structure politique et structures sociales.* Genève: Département de science politique, Université de Genève (thèse).
- Joye, Dominique 1987. Développement méthodologique et analyse du vote. *Annuaire suisse de science politique.* 27. 17-32.
- Kaplan, A. 1964. *The Conduct of Inquiry.* San Francisco: Chandler. chapter 4 and 5.
- King, Gary 1986. How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science.* 30(3) 666-687.
- Kramer, Gerald H. 1983. The Ecological Fallacy Revisited: Aggregate Versus Individual-Level Findings On Economics and Elections, and Sociotropic Voting. *American Political Science Review.* 77. 92-111.
- Kreft, I.G.G., De Leuw, J. and Kim, K.S. 1990. *Comparing Four Different Packages for Hierarchical Linear Regression* Los Angeles: UCLA Statistical Series No. 50.
- Kriesi, Hanspeter (ed.) 1993. *Citoyenneté et démocratie directe. Compétence, participation et décision des citoyens et citoyennes suisses.* Zürich: Editions Seismo.
- Kriesi, Hanspeter 1994. Le défi de la démocratie directe posé par les transformations de l'espace public. in Papadopoulos, Yannis (ed.) *Présent et avenir de la démocratie directe. Actes du colloque de l'Université de Genève.* Genève: Georg. 31-72.
- Kriesi, Hanspeter; Longchamp, Claude; Passy, Florence; Sciarini, Pascal 1993.

- Analyses des votations fédérales du 6 décembre 1992.* Adliswil, Genève: GfS Institut de recherche et Département de science politique, Université de Genève.
- Ladner, Andreas 1993. *Votation populaire du 6 juin 1993 sur les initiatives populaires "pour une Suisse sans nouveaux avions de combat" et "40 places d'armes, ça suffit! - L'armée doit aussi se soumettre à la législation sur la protection de l'environnement". Un résumé de l'analyse des résultats des communes.* Berne: Office fédéral de la statistique.
- Ladner, Andreas 1993. *Aggregatsdatenanalysen: Von Abstimmungsergebnissen in den Gemeinden zum Stimmverhalten der Bürgerinnen und Bürger.* Balstahl: Papier présenté au congrès de l'Association suisse de science politique.
- Ladner, Andreas; Meuli, Urs 1994. *Die Volksabstimmung vom 12. Juni 1994 über das Bundesgesetz vom 18. Juni 1993 über schweizerische Truppen für friedenserhaltende Operationen (Blauhelme).* Zürich: Soziologisches Institut der Universität Zürich.
- Langbein, Laura Irwin; Lichtman, Allan J. 1978. *Ecological Inference.* Beverly Hills: Sage Publications.
- Markus, Gregory B. 1988. The Impact of Personal and National Economic Conditions on the Presidential Vote: A Pooled Cross-sectional Analysis. *American Journal of Political Science.* 32(1) 137-154.
- Meier-Dallach, H.-P.; Nef, Rolf 1987. Values and Identities in Switzerland. *International Political Science Review.* 8(4) 319-332.
- Melich, Anna 1991. *Les valeurs des Suisses.* Berne: Peter Lang.
- Miller, William L. 1977. *Electoral Dynamics.* London: Macmillan.
- Möckli, Silvano *Eine Sekundäranalyse der eidgenössischen Abstimmung vom 6. Dezember 1992 über den EWR-Vertrag.* St.Gallen: Hochschule St.Gallen.
- Nef, Rolf 1979. Bundesabstimmungen 1950-1970. *Bulletin des soziologischen Instituts der Universität Zürich.* 37.
- Nef, Rolf 1980. Struktur, Kultur und Abstimmungsverhalten. *Revue suisse de sociologie.* 6(2) 155-190.
- Nef, Rolf 1988. Rothenturm: das kostenlose grüne "Wunder"? Eine statistische Analyse der Abstimmungen über die Rothenturm-Initiative auf der Basis von 2920 Gemeinden. *Revue suisse de sociologie.* 14(2) 199-224.
- Nef, Rolf 1989. Armeeabschaffungsinitiative: Soyons réalistes, demandons l'impossible? Konzeptionelle Überlegungen und empirische Analysen zum politischen Gehalt einer "Wegmarke" *Revue suisse de sociologie.* 15(3) 545-582.

- Pan, Zhongdang; McLeod, Jack M. 1991. Multilevel Analysis in Mass Communication Research. *Communication research*. 18(2) 140-173.
- Paterson, Lindsay 1991. Multilevel Logistic Regression. in Prosser, Robert, Rasbash Jon and Goldstein, Harvey *Data Analysis with ML3* London: Institute of Education: 5-18
- Paterson, Lindsay 1993. Entry to University by School Leavers. in Woodhouse, G. *A Guide to ML3 for New Users* London: Multilevel Models Project, Institute of Education, University of London.
- Paterson, Lindsay and Goldstein, Harvey 1993. *New Statistical Methods for Analyzing Social Structures: An Introduction to Multilevel Models*. London: Institute of Education.
- Plewis, Ian 1993. Reading progress. in Woodhouse, G. (ed.) *A Guide to ML3 for New Users, 2nd ed..* London: Multilevel Models Project, Institute of Education, University of London:98-124
- Prosser, Bob; Rasbash, Jon; Goldstein, Harvey 1991. *ML3. Software for Three-level Analysis. Users' Guide for V.2.* London: Institute of Education, University of London.
- Prosser, Bob; Rasbash, Jon; Goldstein, Harvey 1991. *Data Analysis with ML3.* London: Institute of Education, University of London.
- Rasbash, Jon 1993. *ML3-E Version 2.3 Manual Supplement*. London: Multilevel Models Project. Institute of Education.
- Roberts, K.H. and Burnstein, L. (eds) 1980. *Issues in Aggregation*. Jossey-Bass, San Fransisco.
- Robinson, W.S. 1950. Ecological Correlations and the Behavior of Individuals. *American Sociological Review*. 15. 351-357.
- Shively, W. Philips 1969. Ecological Inference: The Use of Aggregate Data to Study Individuals. *American Political Science Review*. 63(4) 1183-1196.
- Sprague, John 1982. Is There a Micro Theory Consistent with Contextual Analysis? in Ostrom, Elinor (ed.) *Strategies of Political Inquiry*. Beverly Hills: Sage Publications. 99-122.
- Vatter, Adrian 1994. Der EWR-Entscheid: Kulturelle Identität, rationales Kalkül oder struktureller Kontext? *Revue suisse de sociologie*. 20(1) 15-42.
- Weck-Hannemann, Hannelore 1990. Protectionism in Direct Democracy. *Journal of Institutional and Theoretical Economics*. 146. 389-418.
- Woodhouse, Geoff, Rasbash, Jon, Goldstein, Harvey and Yang, Min 1993. Intro-

duction to multilevel modelling. in Woodhouse, G. *A Guide to ML3 for New Users* London: Multilevel Models Project, Institute of Education, University of London.