JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

G. CHEVRY

Essai de contrôle d'un recensement complet par une méthode d'« Area sampling »

Journal de la société statistique de Paris, tome 93 (1952), p. 56-62

<http://www.numdam.org/item?id=JSFS_1952__93__56_0>

VII

VARIÉTÉ

Essai de contrôle d'un recensement complet par une méthode d' " Area sampling "

Un recensement général de la population fut effectué en France le 10 mars 1946. Comme les précédents, ce recensement comportait en fait plusieurs enquêtes simultanées portant sur les personnes, les ménages, les maisons, les établissements industriels et commerciaux et les exploitations agricoles; des questionnaires différents ont été utilisés pour chacune de ces catégories d'unités statistiques.

Le questionnaire n° 5 concernait les établissements industriels et commerciaux répondant à la définition suivante : un établissement est constitué par un groupe de deux ou plusieurs personnes travaillant en commun d'une manière permanente, en un milieu déterminé, sous la direction d'un ou de plusieurs représentants d'une même raison sociale. D'après cette définition, une personne travaillant absolument seule ne constituait pas un établissement. Deux associés ou bien le mari et la femme travaillant ensemble sans aide formaient un établissement n'occupant aucun salarié. Les diverses succursales d'une même entreprise constituaient chacune un établissement, même si elles étaient situées sur le territoire d'une même commune.

Un premier comptage des questionnaires n° 5 remplis au cours des opérations de recensement permit de constater certaines anomalies qui laissèrent à penser qu'un nombre non négligeable d'établissements n'avaient pas satisfait au recensement. Pour préciser ces impressions et apprécier d'une manière quantitative la valeur du recensement des établissements industriels et commerciaux, la Direction des Enquêtes économiques de l'Institut national de la Statistique décida en octobre 1946 de profiter de la structure régionale de cet Institut pour faire procéder à un certain contrôle *a posteriori* de l'exécution du recensement par une méthode d' « area sampling ».

PRINCIPES DE LA MÉTHODE

Pour réduire le coût de l'opération, le contrôle fut limité aux agglomérations urbaines où se trouve une Direction régionale de l'Institut national de la Statistique. On pouvait d'ailleurs présumer que le recensement du 10 mars 1946 avait été mieux exécuté dans les campagnes que dans les grandes villes.

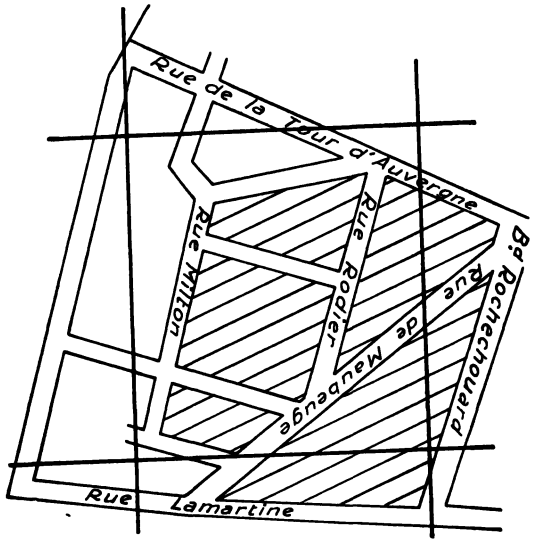
Le contrôle consistait tout simplement à choisir au hasard dans chacune de ces villes un certain nombre de zones où se ferait le sondage, à dresser la liste des établissements situés dans les zones ainsi choisies, à envoyer des enquêteurs visiter toutes ces zones et faire l'inventaire de tous les établisse-

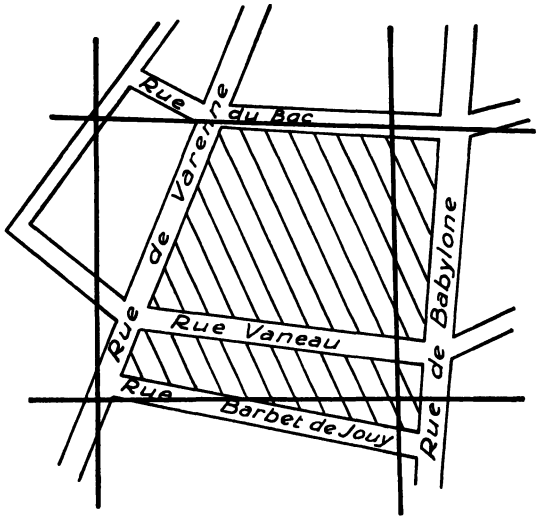
ments qui s'y trouvaient en faisant remplir un questionnaire n° 5 pour les établissements n'ayant pas satisfait au recensement.

Dans chaque ville, il était prescrit de choisir un nombre de zones suffisant pour que le nombre des établissements recensés le 10 mars 1946 dans ces zones soit au moins égal à 2,5 % de l'ensemble des établissements recensés dans l'agglomération urbaine tout entière.

EXÉCUTION PRATIQUE DU SONDAGE

La première opération à réaliser consistait à délimiter les agglomérations urbaines où se ferait le sondage. Il fut prescrit de ne pas le limiter à la ville même où se trouve la Direction régionale mais de le faire porter sur toute





l'agglomération urbaine qui comprend cette ville, cette agglomération étant constituée par l'ensemble des communes ou parties de communes limitrophes dont les parcelles utilisées pour l'habitation sont contiguës ou réunies entre

elles par des parcs, jardins, vergers, chantiers, ateliers ou autres enclos du même genre, alors même que ces habitations ou enclos seraient séparés l'un de l'autre par un fossé, une rivière ou une promenade.

Un plan à grande échelle de l'agglomération urbaine ainsi délimitée fut couvert d'un quadrillage orthogonal régulier dessinant des carrés de 250 mètres de côté pour les agglomérations les moins étendues et de 500 mètres de côté pour les très grandes agglomérations. Les carrés couvrant les zones construites en totalité ou en partie furent numérotés et un certain nombre de carrés tirés au sort au moyen des tables de Tippett.

Comme il était impossible de reconnaître sur le terrain les limites des cases ainsi choisies, il fut prescrit de substituer à chacune de ces cases un polygone de surface sensiblement équivalente et limité par des rues. La figure ci-contre fournit deux exemples de cette substitution qui était indispensable :

a) pour dresser facilement la liste des établissements situés dans chaque case tirée au sort et qui avaient satisfait au recensement.

b) pour désigner sans ambiguïté les immeubles qui devaient recevoir la visite des enquêteurs.

Tous les immeubles situés dans les polygones ainsi choisis et délimités furent visités par des enquêteurs. Chacun d'eux était muni de la liste des établissements recensés le 10 mars 1946 dans le bloc d'immeubles qui lui était imparti. Il pointait sur sa liste les établissements qui avaient satisfait au recensement et faisait remplir un questionnaire n° 5 à ceux qui n'avaient pas été recensés (établissements de deux personnes au moins).

Cette opération ayant eu lieu en novembre 1946, soit plusieurs mois après le recensement général, des instructions spéciales furent données aux enquêteurs pour permettre de distinguer parmi les établissements existant en novembre et n'ayant pas satisfait au recensement, ceux qui, existant déjà en mars, auraient dû être recensés et ceux qui auraient été créés ou réouverts entre mars et novembre. Ces derniers ne sont pas compris dans les résultats présentés ci-après. En revanche, il n'était guère possible de déceler en novembre les établissements qui avaient disparu depuis mars; mais cette lacune paraît sans importance, les cessations de commerce ayant été très rares en cette période d'après guerre.

RÉSULTATS DU SONDAGE

Le tableau I fournit les résultats bruts du sondage pour chacune des villes intéressées et pour l'échantillon tout entier.

Dans l'ensemble des 18 agglomérations urbaines où s'est fait le contrôle par sondage, 201.359 établissements avaient été recensés en mars 1946. L'échantillon choisi a comporté 6.154 établissements recensés soit 3,1 % du total.

L'enquête sur place a révélé dans cet échantillon l'existence de 8.057 établissements. C'est donc que 1.903 établissements, soit 30,9 % des établissements recensés et 23,6 % des établissements existants, y avaient échappé au recensement de mars 1946.

TABLEAU I. — *Résultats bruts du sondage.*

DÉSIGNATION des agglomérations urbaines	NOMBRE total d'établissements (1) recensés le 10 mars 1946 dans ces agglomérations	NOMBRE DES ÉTABLISSEMENTS (1) existant dans l'échantillon prélevé			
		Recensés le 10 mars 1946[1]		Non recensés le 10 mars 1946	
		Nombre	En % du nombre total d'établissements recensés	Nombre	En % du nombre total des établissements dans l'échantillon
1)	(2)	(3)	(4) = (3)/(2)	(5)	(6) = (5)/((3) + (5))
			%		%
Bordeaux	10.556	365	3,4	97	21,0
Clermont	3.185	127	4,0	52	29,1
Dijon	2.366	60	2,5	29	32,6
Lille	7.549	228	3,0	188	45,2
Limoges	2.870	105	3,6	34	24,5
Lyon	14.230	405	2,8	369	47,6
Marseille	14.355	531	3,7	182	25,5
Montpellier	3.281	161	4,9	61	27,5
Nancy	3.272	88	2,7	27	23,5
Nantes	5.005	195	3,9	19	8,9
Orléans	2.748	109	3,9	17	13,5
Paris	110.000	3.115	2,8	529	14,5
Poitiers	905	53	5,9	38	41,8
Reims	3.233	120	3,7	9	7,0
Rennes	2.835	89	3,1	23	20,5
Rouen	5.351	160	3,0	87	35,2
Strasbourg	5.958	128	2,2	60	31,9
Toulouse	4.580	115	2,5	82	41,6
Ensemble	201.359	6.154	3,1	1.903	23,6

(1) Établissements de 2 personnes au moins seuls soumis au recensement.

La théorie de l'échantillonnage permet de déterminer quelle peut être la précision de ce résultat brut. Il ne paraît pas possible d'appliquer purement et simplement la formule qui, dans le cas d'un échantillon en grappes donne la variance de l'estimation d'une probabilité p, cette probabilité étant ici celle pour un établissement d'avoir échappé au recensement. Cette formule suppose en effet essentiellement que toutes les grappes ont la même taille, c'est-à-dire comportent le même nombre d'établissements. Il est évident qu'avec la méthode suivie cette condition n'a pas été remplie.

Dans ces conditions, pour apprécier la précision des résultats, il semble plus indiqué de recourir aux expressions fournies par Goldberg (1) pour l'erreur commise dans l'estimation d'un quotient $\frac{x}{y}$ lorsqu'on utilise le quotient des estimations de son numérateur et de son dénominateur obtenues à partir de n éléments. Cette estimation est entachée d'une erreur systématique dont la valeur moyenne relative est :

$$\frac{C_x^2 - \rho\, C_x\, C_y}{n}$$

(1) Voir *Méthodes statistiques modernes des Administrations fédérales aux États-Unis*, par P. THIONET, p. 58 (Hermann et Cie, Paris, 1946); et *Sampling Theory when the sampling units are of unequal sizes*, par W. G. COCHRAN (Journal of the American statistical Association june 1942).

et d'une erreur-type relative d'échantillonnage :

$$\sqrt{\frac{C_x^2 + C_y^2 - 2\rho C_x C_y}{n}}$$

C_x étant le coefficient de variation des valeurs x, soit : $\frac{\sigma_x}{\bar{x}}$,

$\bar{x}$ la moyenne arithmétique des valeurs x,

C_y le coefficient de variation des valeurs y, soit : $\frac{\sigma y}{\bar{y}}$,

et ρ le coefficient de corrélation linéaire entre x et y.

Dans le cas particulier qui nous intéresse, n est le nombre des polygones limités par des rues tirés au sort, les valeurs y sont les nombres d'établissements existant dans chacun de ces polygones et les valeurs x les nombres d'établissements ayant échappé au recensement dans les mêmes polygones.

Le coefficient de corrélation ρ a été trouvé assez élevé : 0,91. La valeur moyenne relative de l'erreur systématique est $+ \frac{2}{1.000}$ c'est-à-dire pratiquement négligeable. Quant à la valeur relative de l'erreur d'échantillonnage, on a trouvé 7,6 % (1).

On peut donc estimer que le pourcentage des établissements omis par le recensement, par rapport au nombre total des établissements existants, a 95 chances sur 100 d'être compris

entre 23,6 $(1 - 2 \times 0{,}076)$ soit 20 %
et 23,6 $(1 + 2 \times 0{,}076)$ soit 27,2 %.

et affirmer que le recensement du 10 mars 1946 a laissé échapper dans les grandes villes françaises de 20 à 27 % des établissements industriels et commerciaux.

On a déjà signalé plus haut qu'il ne paraît pas *a priori* légitime d'étendre ce résultat à la France tout entière. Le sondage n'a été effectué que dans les grandes villes et il n'y a aucune raison valable de supposer que le dénombrement des établissements a été aussi incomplet dans les campagnes que dans les villes.

RÉSULTATS COMPLÉMENTAIRES

S'il est fort intéressant de savoir que, dans les grandes villes, de 20 à 27 % des établissements n'ont pas été touchés par le recensement général, ce renseignement est toutefois insuffisant. Il convient de se demander comment et pourquoi le recensement a été incomplet. En particulier, la première question qui vient à l'esprit est la suivante : Est-ce que le fait d'avoir échappé au recensement dépend de la nature et de la taille des établissements? On pourrait en effet penser que ce sont surtout de petits établissements qui ont été omis.

---

(1) Ces calculs n'ont pu pratiquement être conduits que pour 14 des 18 villes où avait eu lieu le sondage. Pour les quatre autres, les résultats relatifs à chacun des polygones tirés au sort n'ont pas pu être exploités. Cette circonstance n'infirme en rien les conclusions que l'on peut tirer des résultats obtenus, les marges d'erreur calculées sur 14 villes seulement étant plus grandes que celles qui auraient été trouvées pour les 18 villes.

Une réponse à cette question, du moins en ce qui concerne la taille des établissements, est fournie par le tableau II qui compare la répartition suivant l'effectif de leur personnel salarié de 756.235 établissements recensés le 10 mars 1946 dans la France entière à la répartition analogue de 1.623 établissements révélés par le sondage comme ayant échappé au recensement.

Les deux séries de pourcentages marquent une concordance extrêmement satisfaisante pour tous les établissements comptant 2 salariés et plus. Pour les deux autres catégories (0 et 1 salarié) la concordance est moins bonne. Cependant il ne semble pas qu'il faille y attacher trop d'importance. Tout d'abord, les pourcentages résultant du sondage ne peuvent être tenus que pour approximatifs en raison des erreurs d'échantillonnage. En outre, le classement, dans l'une ou l'autre de ces deux catégories, des établissements qui leur correspondent présente en fait une certaine part d'incertitude. Il ne faut pas oublier en effet que le recensement de 1946 et le sondage ne portaient que sur les établissements de deux personnes au moins. Ceux d'entre eux qui groupaient deux personnes seulement pouvaient être classés soit comme comportant deux patrons et par suite 0 salarié, soit comme comprenant un patron et un salarié, suivant l'interprétation donnée au questionnaire par le déclarant. Il est d'ailleurs fort possible que, notamment pour des raisons d'ordre fiscal, un établissement ayant deux patrons et 0 salarié ait déclaré avoir un patron et un salarié c'est-à-dire que la femme du patron ait été indiquée à tort comme salariée.

TABLEAU II. — *Répartition suivant l'effectif de leur personnel salarié de 756.235 établissements recensés en mars 1946 et de 1.623 établissements non recensés mais révélés par le sondage.*

EFFECTIF DU PERSONNEL salarié	ÉTABLISSEMENTS RECENSÉS		ÉTABLISSEMENTS NON RECENSÉS révélés par le sondage	
	Nombre	% du total	Nombre	% du total
0 salarié	230.354	30	397	25
1 salarié	197.252	26	546	33
2 à 5 salariés	214.504	29	443	27
6 à 10 —	48.561	6	100	6
11 à 20 —	29.188	4	65	4
21 à 50 —	21.294	3	47	3
51 à 100 —	8.001	1	15	1
plus de 100 salariés	7.081	1	10	1
Ensemble	756.235	100	1.623	100

Dans ces conditions, il paraît prudent de confondre en une seule les deux premières catégories du Tableau II. On obtient ainsi des pourcentages de 56 % au recensement et de 58 % au sondage, ce qui peut être considéré comme suffisamment concordant.

Du reste, si l'on applique le test χ^2 de Pearson aux deux répartitions ci-dessus, en réduisant à cinq le nombre des catégories pour que chacune comporte un nombre d'observations suffisant, on trouve $\chi^2 = 2{,}02$. Les tables de probabilité de χ^2 indiquent, pour quatre degrés de liberté, qu'il y a plus de 70 chances sur 100 pour que χ^2 dépasse 2. Ce résultat signifie que les établisse-

ments ayant échappé au recensement peuvent être considérés comme ayant été choisis absolument au hasard parmi les établissements existants.

Il semble donc possible de conclure que si le recensement de 1946 a été incomplet, il l'a été pour les établissements de toutes tailles. D'après les indications fournies par les enquêteurs du sondage, c'est à la négligence des agents recenseurs qu'il faut attribuer cette imperfection du Recensement.

Ces constatations ont conduit l'Institut National de la Statistique à renoncer à publier tels qu'ils se présentaient, les résultats du Recensement 1946 concernant les entreprises industrielles et commerciales.

Depuis, des recherches ont été faites à ce sujet par d'autres moyens. Elles ont conduit à la constitution d'un inventaire permanent des entreprises qui donnera une physionomie beaucoup plus exacte et complète de la structure industrielle et commerciale de la France (1).

G. Chevry.