PROGRAM OF RESEARCH AND ACTIONS ON THE DEVELOPMENT OF THE LABOUR MARKET

ANALYSIS OF THE RESULTS OF THE LABOUR FORCE SAMPLE SURVEYS

5. STRUCTURE BY AGE GROUP OF MALE EMPLOYMENT AND FEMALE EMPLOYMENT BY ECONOMIC ACTIVITY AND OCCUPATIONAL STATUS

(1977 DATA)

By :

Jacques A. ZIGHERA
Annick GUILLOUX
Nicolas BOUSCARLE
Centre d'Analyse Statistique
des Structures et des Flux
Université de Paris X Nanterre

For :

The Commission of the
European Communities

PROGRAM OF RESEARCH AND ACTIONS ON THE DEVELOPMENT
OF THE LABOUR MARKET

ANALYSIS OF THE RESULTS OF THE LABOUR FORCE
SAMPLE SURVEYS

5. STRUCTURE BY AGE GROUP OF MALE EMPLOYMENT AND FEMALE EMPLOYMENT BY ECONOMIC ACTIVITY AND OCCUPATIONAL STATUS

(1977 DATA)

By :

Jacques A. ZIGHERA
Annick GUILLOUX
Nicolas BOUSCARLE
Centre d'Analyse Statistique
des Structures et des Flux
Université de Paris X Nanterre

For :

The Commission of the
European Communities

SUMMARY

The authors attempt to compare the employment profiles by age group and by economic activity and occupational status in the various countries of the Community. Since the employment distribution by age group as well as its distribution by activity differs from one country to another, the authors establish, on the basis of Information Theory, « profiles » by age group in each activity in the nine countries ; these profiles, which indicate the relative over- or under-representation of the age groups, allow comparisons of different countries to be made.

RÉSUMÉ

On désire comparer les profils par classes d'âges de l'emploi par branches (et statuts), dans les divers pays de la Communauté. La répartition par classes d'âges de l'emploi étant différente d'un pays à l'autre, de même que la répartition par branche, on calcule, en s'appuyant sur la théorie de l'information, des « profils » par classes d'âges pour chaque branche dans les neuf pays ; ces profils qui donnent la sur- ou sous-représentation relative des classes d'âges permettent d'effectuer des comparaisons entre pays.

INTRODUCTION

0.1. In preceding studies[1] or studies in progress[2] we have examined the activity rates or employment rates (i.e. not including the unemployed in the case of the latter) according to age in the various member countries of the European Community, and for men and women. It is likely a priori that, if one examines employment by breaking it down by activity and occupational status, one will find perceptible differences in profiles by age ; thus, one can expect to find, all other things being equal, few young men among the self-employed and few aged men in industries where retirement occurs very early. The aim of this report is to study these problems thoroughly by analysing the data in this field that are available in the labour force sample surveys.

0.2. Such a study can be conceived of on different levels ; it is fitting, therefore, to specify the limits of this report, which will be transcended in later stages. First and foremost, the analysis is static and treats a single year, 1977 ; it will, of course, be very useful in the future to analyse the evolution, which will make possible the study of phenomena such as entries and exits (or rather the balance of entries and exits) in the different activities and for the different ages ; but since the data now available and used here concern age groups of five years, this more thorough study will be carried out later, when we can follow a generation of a given year of birth through several sample surveys. Next, we shall introduce no regional differentiation. Last, the industries chosen are the eleven branches of economic activity of the NACE with subdivision into self-employed, manual workers and non manual workers, for agriculture and industry, and into self-employed and employees for the services. We shall use the term « activity-status » to designate these groups.

(1)   same series, instalment 2 (out of print)
(2)   same series, instalment 6.

0.3. As the following plan shows, this report analyses, on the one hand, the age structure of the male population, and on the other hand, that of the female population ; there is no attempt to analyse here the more or less great partici pation of women in different activities.

0.4. Details of the method used are given in the appendix. It is enough to indicate here that since the aim is to compare employment structures cross-classified by age group and by activity-status, between countries where employment structures by age group are different as well as employment structures by activity-status, one must therefore find a means of rendering these cross-classified employment structures comparable, i.e. independent of national distributions by age group and by activity. Thus, it will not be necessary to insist each time, for example, on the fact that the employment of youth is lower in Italy or that employment in the mining and chemical industry is relatively higher in Luxemburg, but one should be prepared to compare relative situations.

Starting from Information Theory, one is led to calculate coefficients equal to (for a given country) :

$$\frac{\text{(employment in an activity-status, in an age group)} \times \text{(total employment)}}{\text{(employment in the activity-status)} \times \text{(employment in the age group)}}$$

These coefficients are graphically represented on a logarithmic scale ; this is what we shall call profiles. Moreover, the weighted average of the logarithms of these coefficients gives us, for a given country, mutual information, i.e. a measure of the extent of the differences ; thus one will find less mutual information for the United Kingdom than for F.R. Germany, this meaning that the differences of profile by age between activity-status are less marked in the former country than in the latter. Finally, on the basis of these profiles one can establish a typology of activity-status (this too is treated at greater lengh in the appendix).

0.5. As indicated above, the data used here come from the 1977 labour force sample survey. We have, for each of the member countries and for the Community as a whole, tables of the labour force cross-classifying the activity-status with twelve age groups.

0.6. In the different tables and texts that follow we have used these classifications and symbols (in this order) :

— for the occupational status :

I	:	employers and self-employed
O	:	manual workers
E	:	non-manual workers
S	:	employees

(for the services, the notable differences between countries in the relative importance of manual and non-manual lead us to think that the criteria for distribution between these two categories are not the same, and that an analysis carried out on this basis would lead to erroneous conclusions).

— for the activities :

a	:	agriculture, forestry, fishing
E	:	energy and water
X	:	extraction and processing of non-energy-producing minerals and derived products ; chemical industry
I	:	metal manufacture ; mechanical, electrical and instrument engineering
M	:	other manufacturing industries
B	:	building and civil engineering
C	:	distributive trades, hotels, catering, repairs
T	:	transport and communication
A	:	banking and finance, insurance, business services, renting
G	:	public administration
S	:	other services.

(in the body of the text, we shall use shortened expressions such as agriculture or extractive and chemical industry, but they will always correspond to the complete definitions given above).

0.7. On the graphs, we shall use the usual colours for each country, i.e. :

	solid	dashes
blue	France	
yellow	F.R. Germany	
brown	Belgium	Luxemburg
red	Netherlands	Denmark
green	Italy	Ireland
violet	United Kingdom	

0.8. To make it unnecessary to repeat ourselves in each of the following parts, it should be kept in mind that the data come from a sample survey that is indeed large, but by cross-classifying activities and age groups, it will happen that certain results are not fully significant, especially in the activities employing a small share of the labour force in the countries other than the four most populous.

1. ANALYSIS OF THE AGE DISTRIBUTION OF THE
EMPLOYED MALE POPULATION

1.1. One should clarify first of all a point of vocabulary linked to the method used. When we write that in a given country there are many men from 45 to 49 years of age in a certain activity, one should always have in mind that we mean that, once taken into account the size of the employment in the activity in this country and the number of employed males 45 to 49 years of age in the country, there are relatively a great many men 45 to 49 in this activity. This approach will precisely enable us to compare the structures of countries where the sizes of activities and age groups are different.

1.2. When one engages in the analysis that is briefly described in the introduction and amplified in the appendix, one observes certain general results. In what concerns mutual information, which measures the importance for a given country of difference in age structure of the activities (or vice-versa), one obtains the following numerical results (measured in thousandths), with the countries arranged in decreasing order :

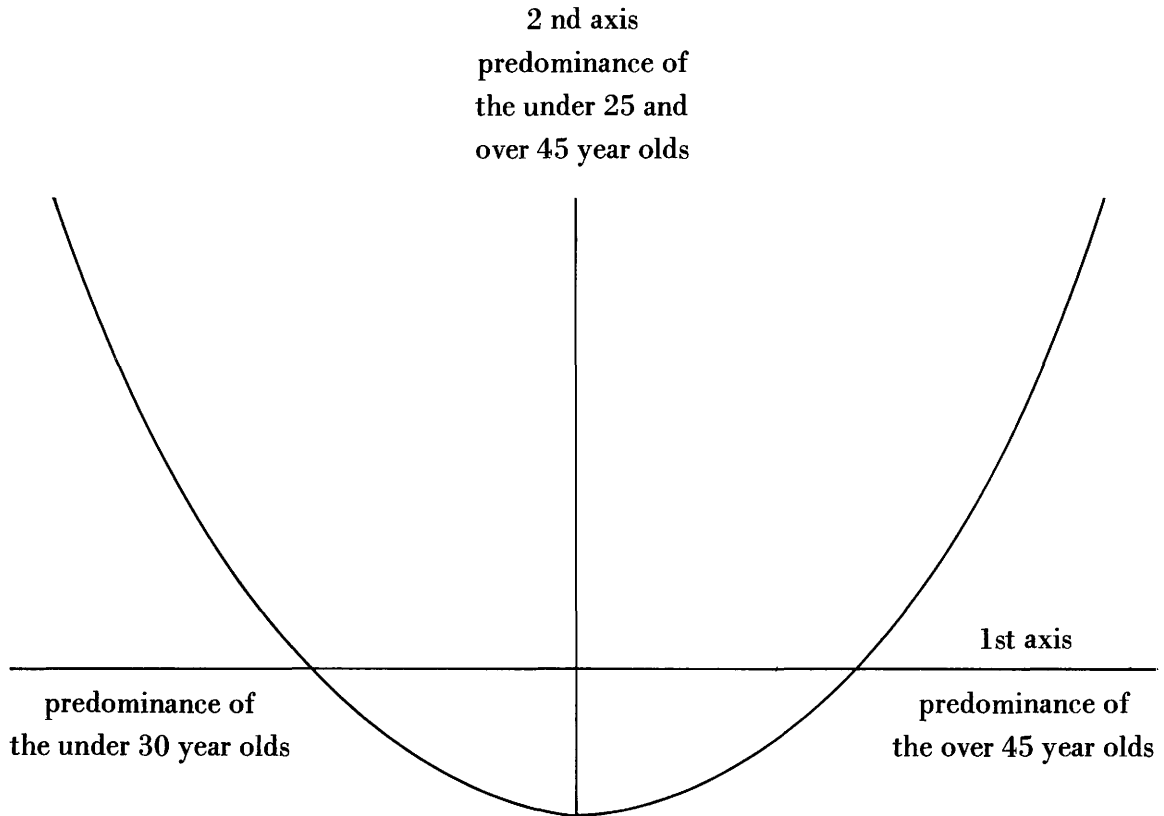|	
Ireland	169
Luxemburg	167
Denmark	141
Italy	118
France	105
F.R. Germany	96
Belgium	84
Netherlands	68
United Kingdom	51

One sees in this table that the United Kingdom and the Netherlands are the countries with the slightest differences in their structures and Luxemburg and Ireland the countries with the greatest differences. One should nonetheless note that mutual

information includes a part originating in the sampling error, and that one can compare directly only the values corresponding to countries where the samples are of the same size.

1.3. If one now proceeds to analyse the data according to the factors extracted from the tables, the contribution of the different axes is the following :

1st axis	53 %
2nd axis	22 %
3rd axis	9 %
4th axis	6 %

One sees immediately that the first axis opposes the size of the under-30 group to that of the over-45 group, and that the second axis opposes the size of the 25-45 group to the size of both the youngest and the oldest groups. We shall therefore use a general diagram that is easy to analyse in the plane of the first and second axes (see figure below).

2 nd axis
predominance of
the under 25 and
over 45 year olds

1st axis

predominance of
the under 30 year olds

predominance of
the over 45 year olds

1.4. On the graph on this page, in the case of the Community as a whole, we have represented the projection, on the first two axes, of the results of the data analysis concerning the profiles of activity-status. To facilitate the reading of the graph, employees in the services are shown in blue, manual and non-manual workers in industry and agriculture in red and violet respectively, and self-employed in green. Referring to the preceding graph, one observes the relative over-representation of the young among employees in credit and insurance, and still more in distributive trade, and among the manual workers of the metallurgical-mechanical and of the other-manufacturing and building industries, and the relative over-representation of the oldest among the self-employed, especially in agriculture.

1.5. The profile of the various groups are shown farther on, but as an illustration we shall now show on the graph on the following page some characteristic profiles for the entire Community. One thus observes that :
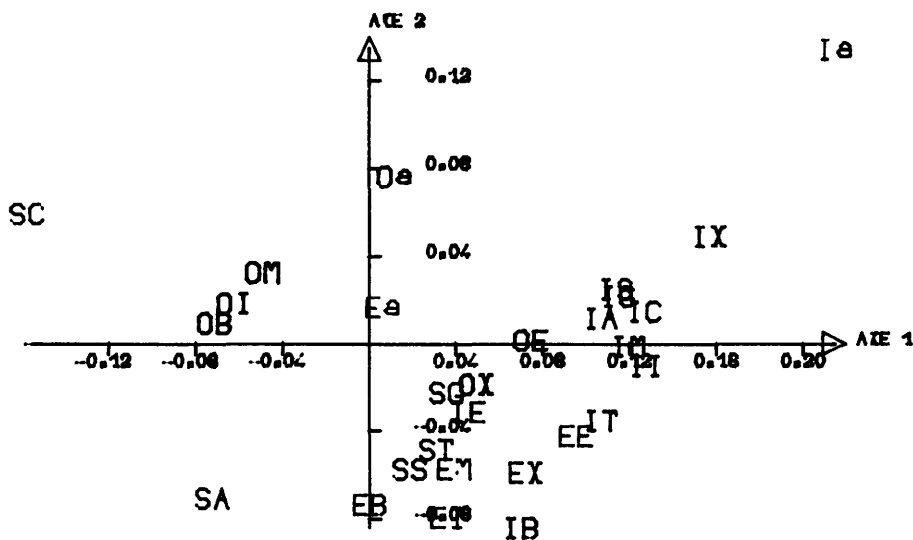
— in the case of self-employed in agriculture (green line) aged persons are greatly over-represented and the young under-represented ; one observes that there are twice as many self-employed of agriculture in the 60-64 age group as in the overall working population and half as many in the 30-34 age group ;

— in the manufacturing industry, up to 25 years of age, manual workers (red dashes) are over-represented, and non-manual workers (violet) are under-represented ; for manual workers over 65, under-representation becomes great (probably due to more systematic retirement).

— for manual workers in the extractive and chemical industry (solid red), the profile is differentiated from that of the manufacturing industry by the absence of over-representation of the young and a still greater under-representation starting at age 65 ;

— finally, if one considers employees in distributive trade (solid blue) and insurance-credit (blue dashes), one observes a great over-representation of the young (except for those under 20 in insurance-credit, this corresponding to the fact that distributive trade recruits young people who have left school earlier) and an under-representation of the middle age groups (40 to 60 years).
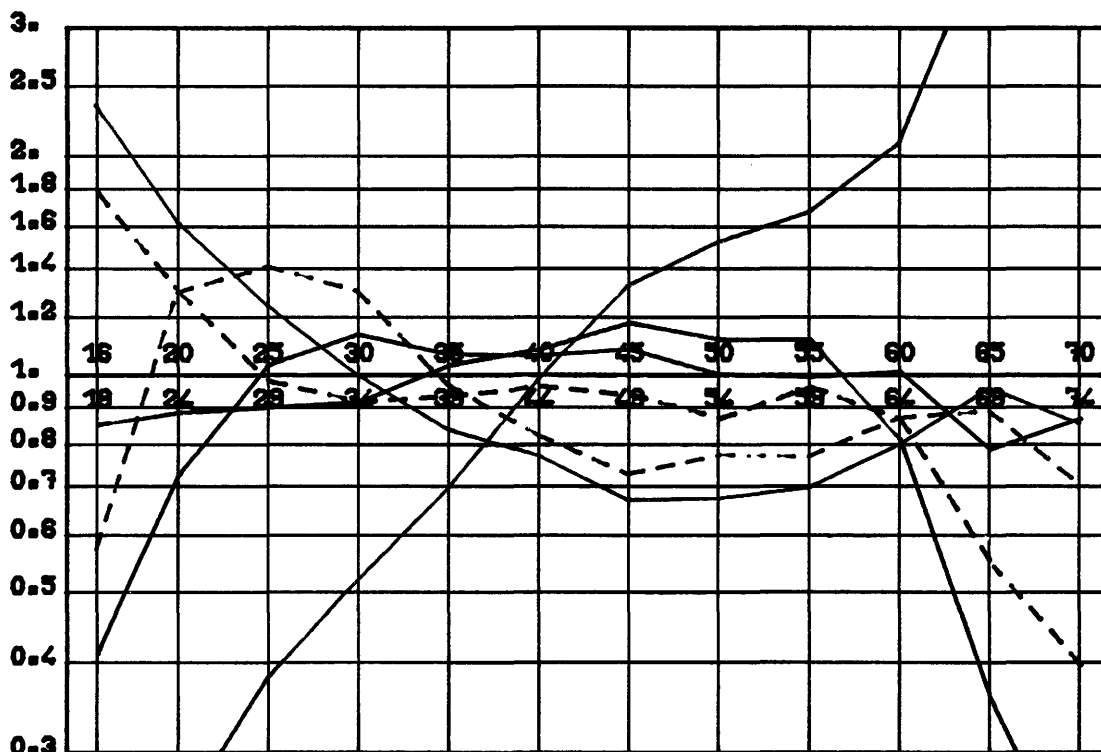
1.6. We shall now present the different profiles, comparing each time the profiles of the different countries for the same activity-status. In some cases we have regrouped certain closely related categories and in others, where the labour force is very small, the graphs of the profiles are not shown.

1.7. The *employees in distributive trade* constitute the first, very characteristic group with, as we have already seen, a very great over-representation of the very young. In all the countries, persons under 30 are over-represented, and those in the 16-19 age group are, proportionately at least, twice as numerous in this group as in the total employment. For the very young, the phenomenon is particularly marked in Italy and Luxemburg and, to a lesser extent, in Ireland. One also observes that in what concerns the 25-29 year olds, over-representation is distinctly greater in Denmark, Italy, and F.R. Germany than in the other countries. Inversely and in compensation, the under-representation of those over 40 is marked, especially in Italy, F.R. Germany, Denmark, and Luxemburg. Just under or over the age of 60, there seems to be a certain
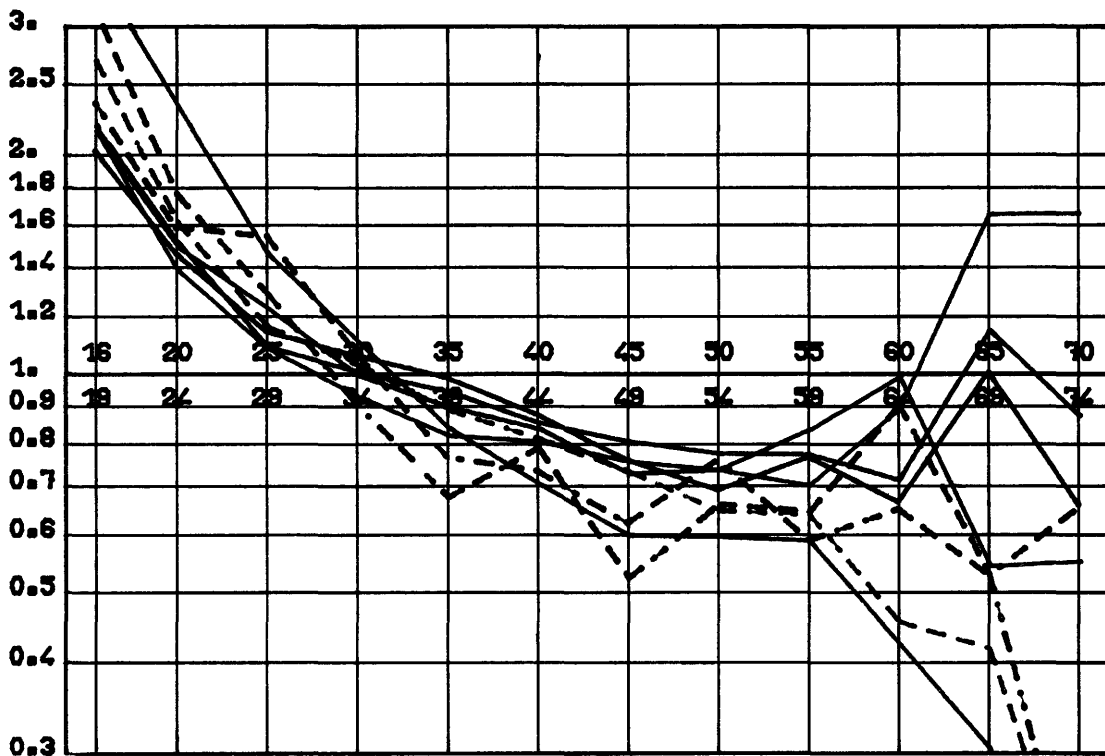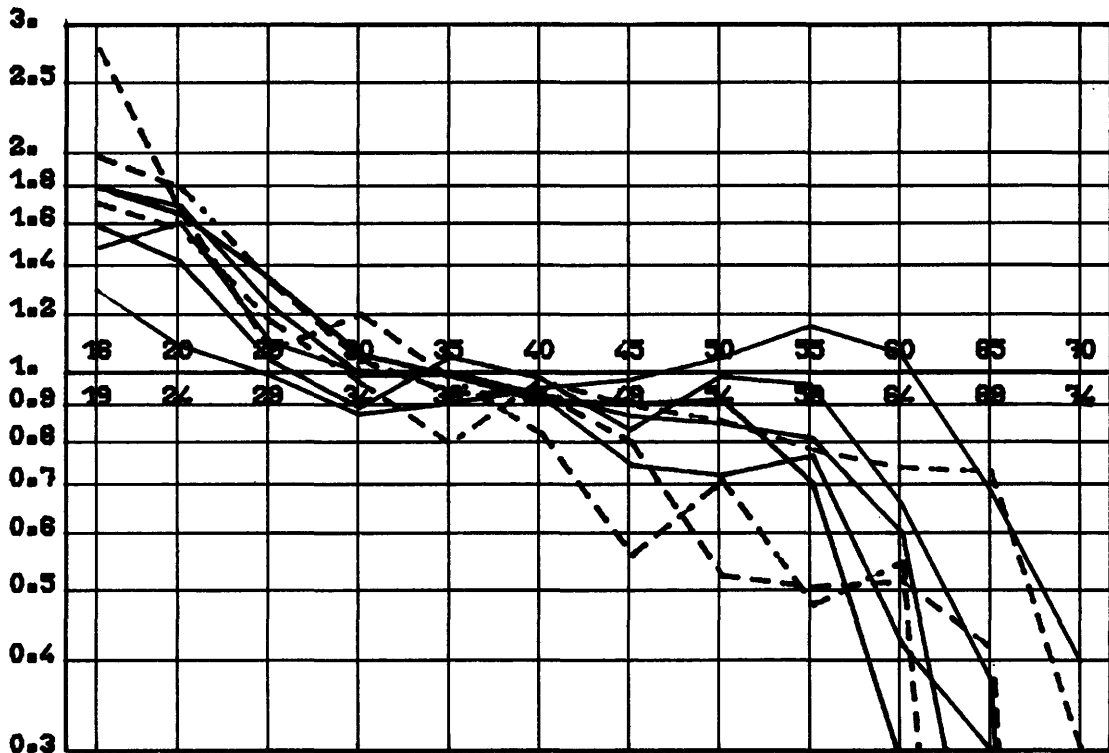
over-representation (or less under-representation) in certain countries : the United Kingdom, France, Belgium, and the Netherlands. Here as elsewhere (and perhaps more here than elsewhere) the profiles thus obtained are very similar for the various member countries.

1.8. One should observe that *insurance and finance employees* are much less numerous than those in distributive trade and that the profiles shown below are therefore based on smaller and therefore less precise samples than the preceding one, especially for the oldest age groups where the number of persons employed is particularly low. The general profile is distinguished from the preceding one by a sort of shifting : the very young are under-represented, but inversely, over-representation is the rule for those under 35 ; the under-representation of the very young corresponds probably to the fact that employment in this activity requires, on an average, a higher education than in the preceding case. If one now examines the profiles of the different countries, one observes that in what concerns the youngest age group, from 16 to 19 years, the
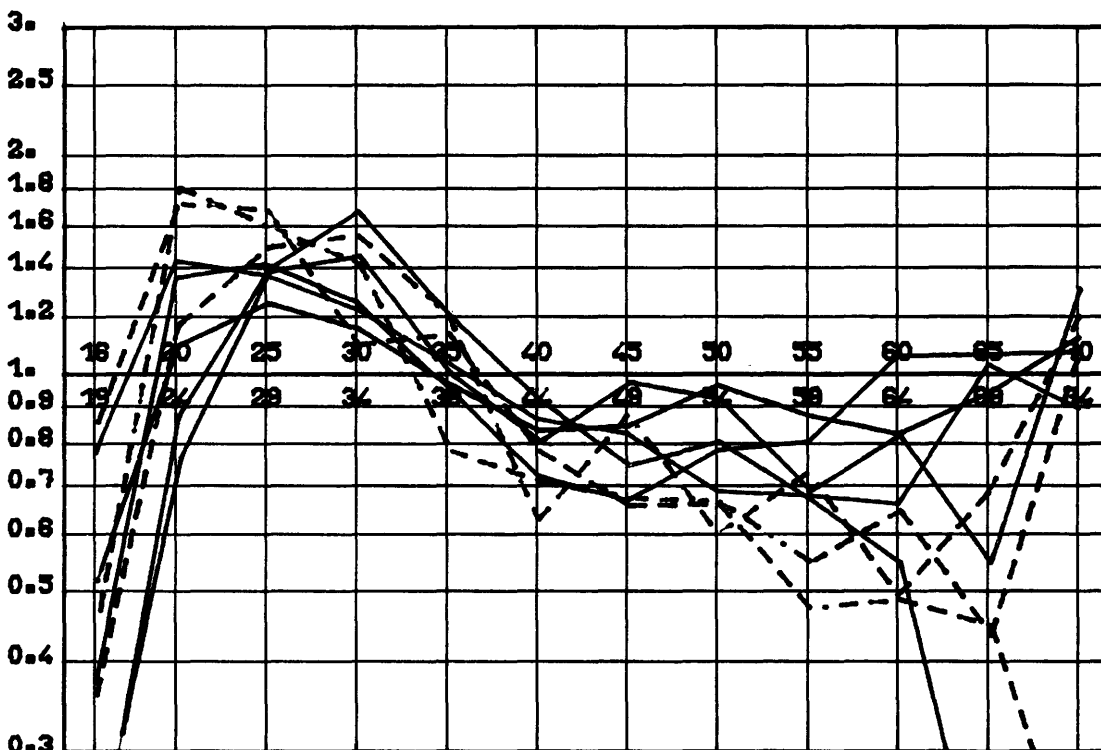
countries can be divided into three main categories : first, Ireland and the United Kingdom where under-representation is slight, then Belgium and Italy where it is high (about a quarter of the average level), and finally the other countries (about one-half). At the ages between 20 and 35 years one can distinguish between Ireland and Luxemburg, with a high over-representation of the 20-30 year olds, and Belgium, Italy, and Denmark with over-representation of the 25-35 year olds. Although one cannot now be certain, since only dynamic analysis will permit a more accurate definition, it is likely that in these countries this activity recruited a great deal several years ago, bringing about a low in more recent recruitments. Finally, the general under-representation of persons over 40 is particularly marked for Luxemburg, Ireland, and Denmark.

1.9. *Manual workers in the metallurgical and mechanical industries and in the manufacturing and construction industries* have relatively similar profiles, i.e. in

most countries over-representation until 25 or 30 and rather distinct under-representation for older age-groups. While doubts are often expressed as to the capacity of the secondary sector to furnish employment, it is fitting to stress the fact that these graphs indicate that, for most of the countries, there are in the 16-19 age group, 1.5 to 2.5 times as many workers in these sectors as in the total of employed persons ; there again, only a dynamic analysis will enable us to obtain a detailed picture of the evolutions, but it is fitting to stress the fact that these groups furnish therefore a relatively high number of jobs to those who start to work young. Inversely, one must moderate this commentary by stressing that further along (1.13.) employees of the manufacturing industries are under-represented in the young age groups and that this may in part indicate shifts in occupational status that will come with age.

1.10. If one examines more especially the *manual workers in the metallurgical and mechanical industries* (see graph on preceding page), one observes a very homogeneity

in the profiles : over-representation (coefficient of about 1.6.) from 16 to 24 years of age, average profile until about 55-60 years of age, and then under-representation corresponding to retirements ; there are two exceptions, one very distinct one in the United Kingdom whose profile is very flat, with a very slight over-representation of the young and, inversely, under-representation that does not continue beyond the age of 65 ; the other one, less distinct, concerns Luxemburg where over-representation of the 16-19 year olds and under-representation beyond the age of 45 is more marked. For *manual workers in manufacturing industries* (see graph opposite), except in Italy, under-representation is distinct from the age of 50 on, while, among the youngest, one observes the very high over-representation of the 16-19 year olds (coefficient of 2.5), in France, Belgium, and the Netherlands ; inversely, if one considers all persons under 25, over-representation is less marked in Germany than elsewhere ; there is also a less marked under-representation between 60 and 70 in the United Kingdom, Denmark, and Ireland. Finally, in the case of *manual workers in building industry* (see graph below) the over-representation of the youngest is particularly marked in (in this order) Belgium and Italy, Ireland and Luxemburg, and France.
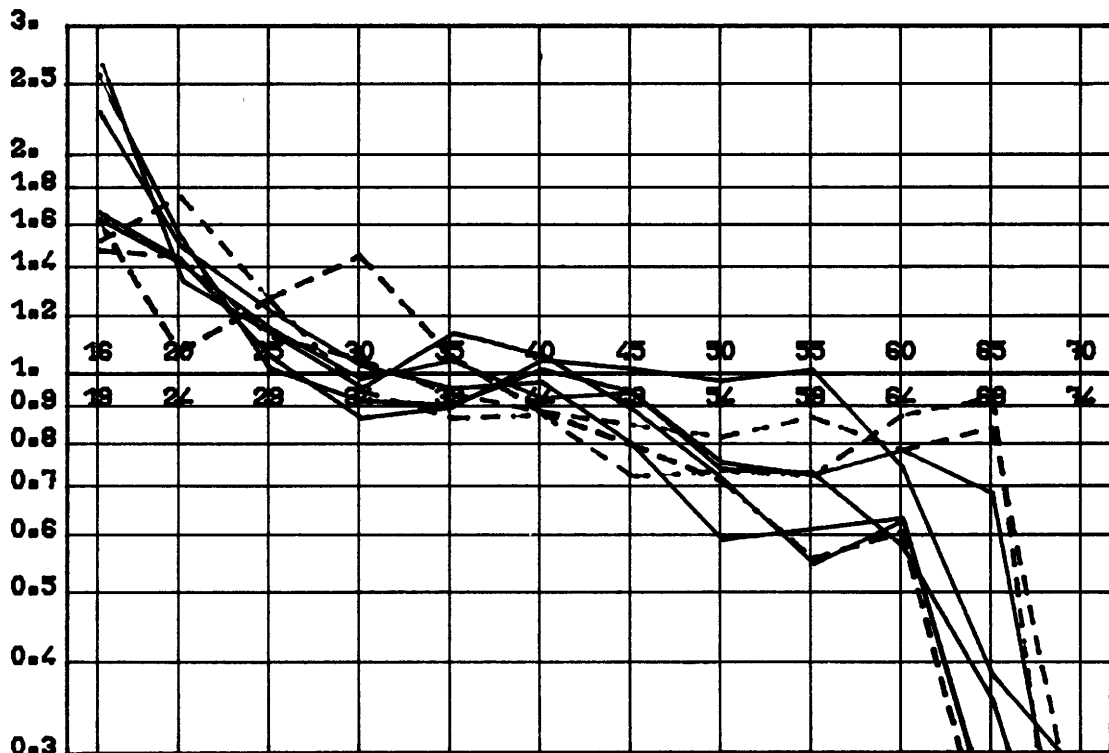
1.11. The *employees in remaining services* i.e., transports, general administration and other services, are characterized by rather flat profiles, corresponding to the fact that on the graph on page 11 the representative points are rather close to the origin. They have in common the under-representation of persons under 20 and in certain cases under 25, then a rather flat profile beyond. A closer examination of *employees in transports* (see graph above) shows that the under-representation of persons under 20 is pronounced in France and Belgium, and that of persons under 25 in the United Kingdom and even more so in Italy ; under-representation in the older age groups becomes pronounced at 60 in Luxemburg, Italy, France, and a little less so in Denmark ; it is not at all pronounced in the United Kingdom. In the case of *general administration employees* (see graph opposite, above), there are, first of all, several countries having a profile with two maxima of over-representation, around 20-30 years on the one hand, and around 50-60 years on the other ; this phenomenon is particularly pronounced in the Federal Republic of Germany, Denmark, and the Netherlands ; under-representation begins in most countries only around the age of 65. Above all, one should note the great under-representation of persons under 25 in France
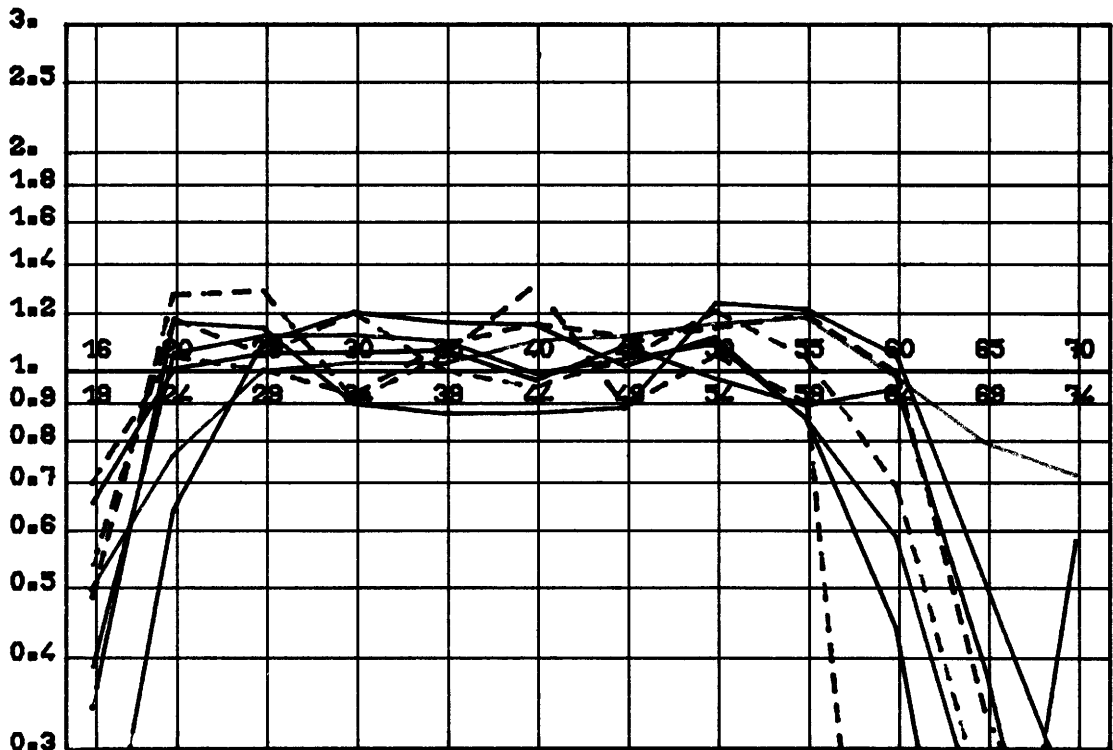
and of those under 30 in Italy ; it is striking to see among the 20-24 year olds the amplitude of differences among countries, with Germany, Denmark, and Ireland at one extreme (coefficient of about 1.4) and France (0.7) and especially Italy (0.4) at the other. There again, a dynamic study proves necessary in order to separate what can possibly be explained by the different educational level required and by the conse= quences of budgetary restrictions in the domain of job creation. Finally, concerning the *employees of other services* (see graph on preceding page, below), one observes a certain over-representation of the 25-35 year olds and the absence of under-representation before the age of 70 ; the under-representation of persons under 25 is particularly pronounced in F.R. Germany and Denmark. The profiles of the United Kingdom and Italy are distinctly flatter, and there is also a certain over-representation beyond the age of 50.

1.12. *Manual workers in the extractive and chemical industry and in energy* (see graph below) were situated much more to the right than other groups of workers on the graph on page 11. On the corresponding profiles there are very distinct characteristics
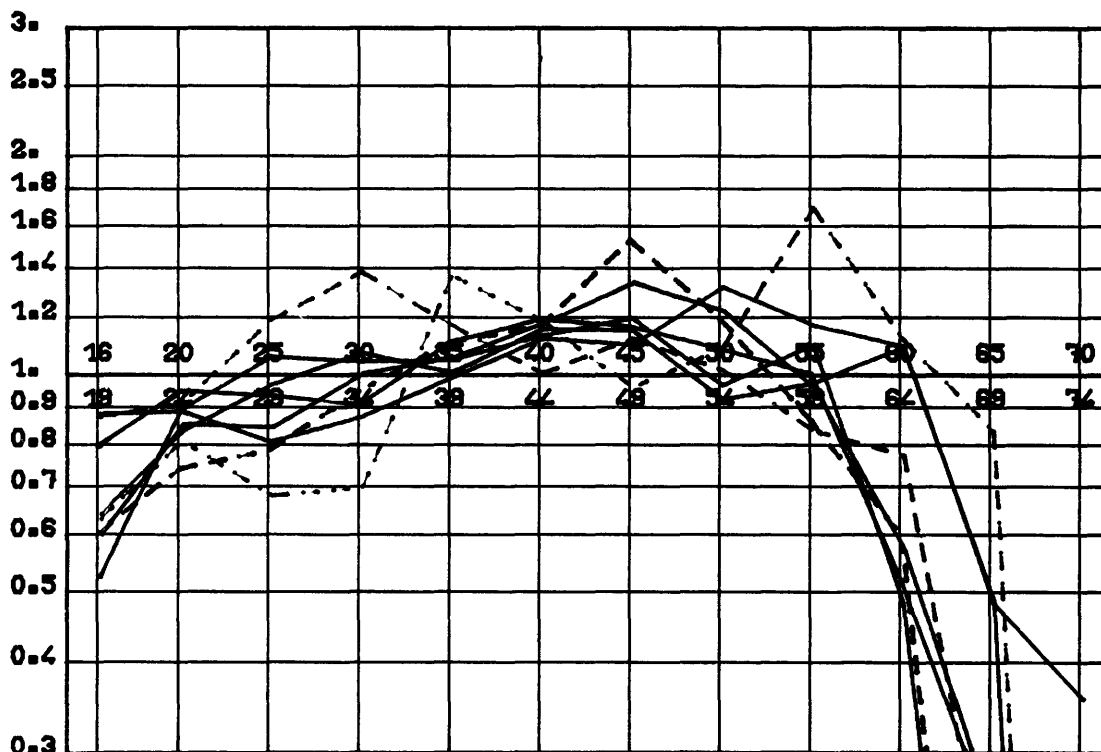
that are common to almost all the countries : profile rising very slightly towards the left with a progressive transition from under-representation among the young to over-representation culminating around the age of 50 or a little under, and then, for most countries, a very rapid decline of the profile around the age of 60. A more precise examination by country indicates that the under-representation of the very young is particularly pronounced in Italy, Luxemburg, France, the Netherlands, and Denmark ; the precision is linked to the portion of the number of workers concerned, and it is particularly small in the three smallest countries, which explains various irregularities. Of course, even though one does not dispose of figures, it appears that in most countries the profile is dominated by the manual workers of the extractive industry, there being little recruitement and rather early retirement.

1.13. *Non-manual workers in industry* (below) constitute only 10 % of the employed, and an examination of the graph on page 11 shows them to be rather similar in all activities. There is, first of all, a very great under-representation of the 16-19 year olds, around
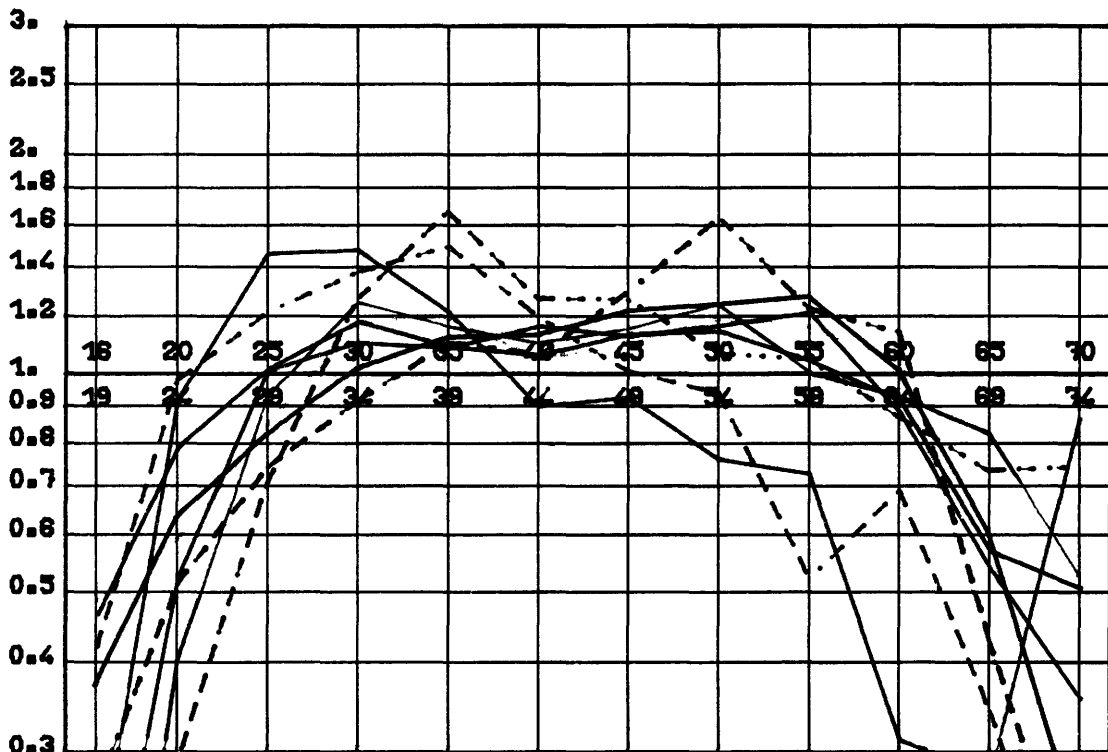
half for the United Kingdom, F.R. Germany, Ireland, and Denmark, and still less for the others. The fact that this under-representation ceases in general only around the age of 25 or 30 leads one to think that it corresponds in part to the shifting of categories ; most profiles are then rather flat up to the age of 60. One should note the particular profiles of Ireland and Italy where the 25-40 age group is greatly over-represented.

1.14. Finally, leaving aside the manual workers and non-manual workers in agriculture, of whom there are very few, especially in certain countries, employers and the self-employed remain to be examined ; we have already seen that the representative points of the latter correspond to profiles dominated by the higher ages, this being particularly pronounced in the case of employers and the self-employed in agriculture and much less so in the building industry ; we shall examine successively these two groups and then the group of employers and the self-employed in distributive trades where the differences among countries are particularly pronounced, and finally the remaining employers and self-employed. In what concerns *employers and the self-employed in agriculture* (see graph below), the profile is very sharp and very regular for all countries except the United Kingdom ; for this country under-representation of the young and over-representation of
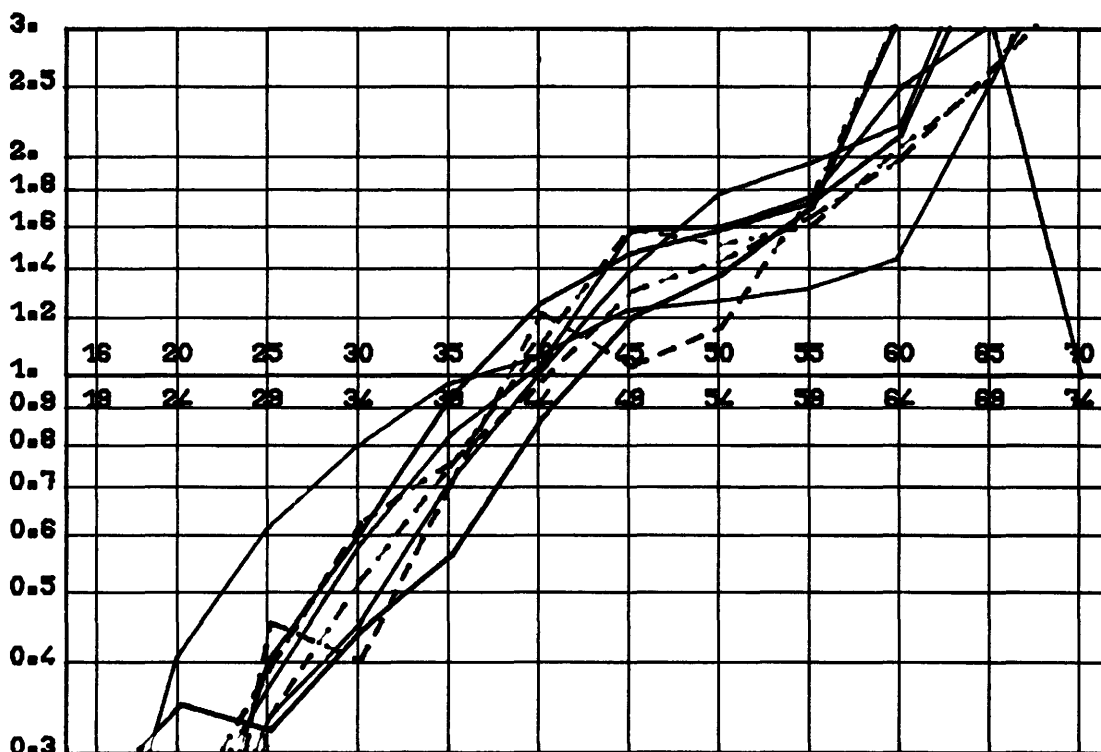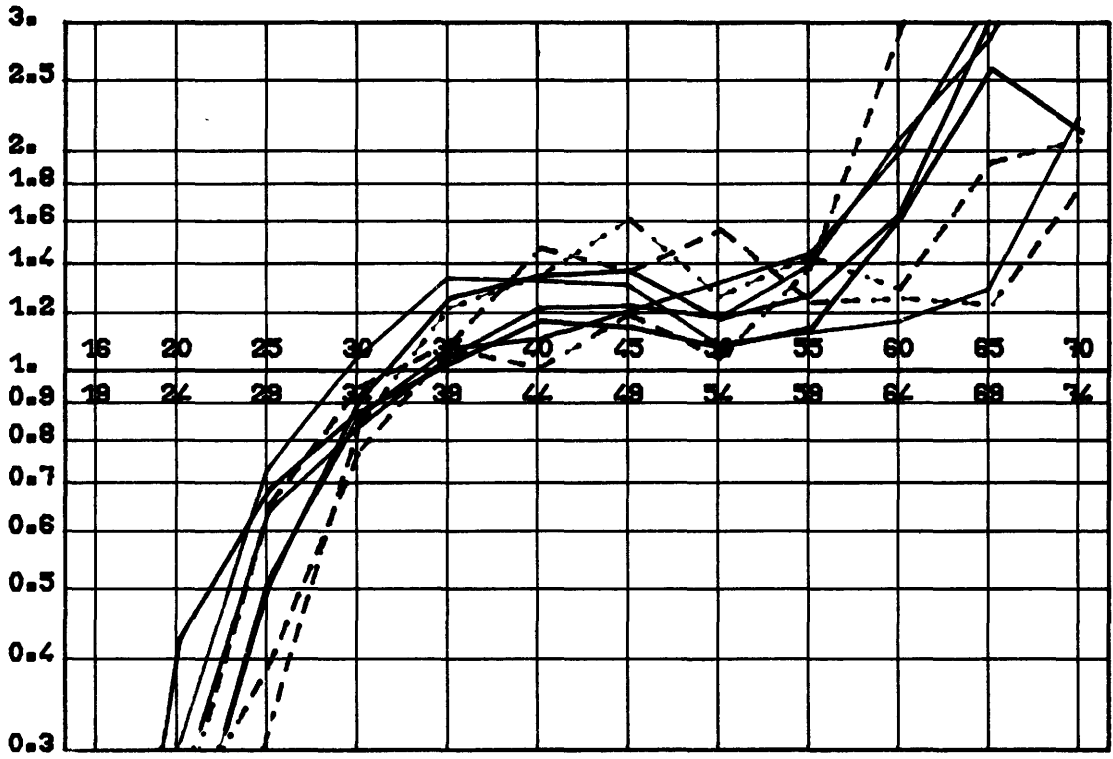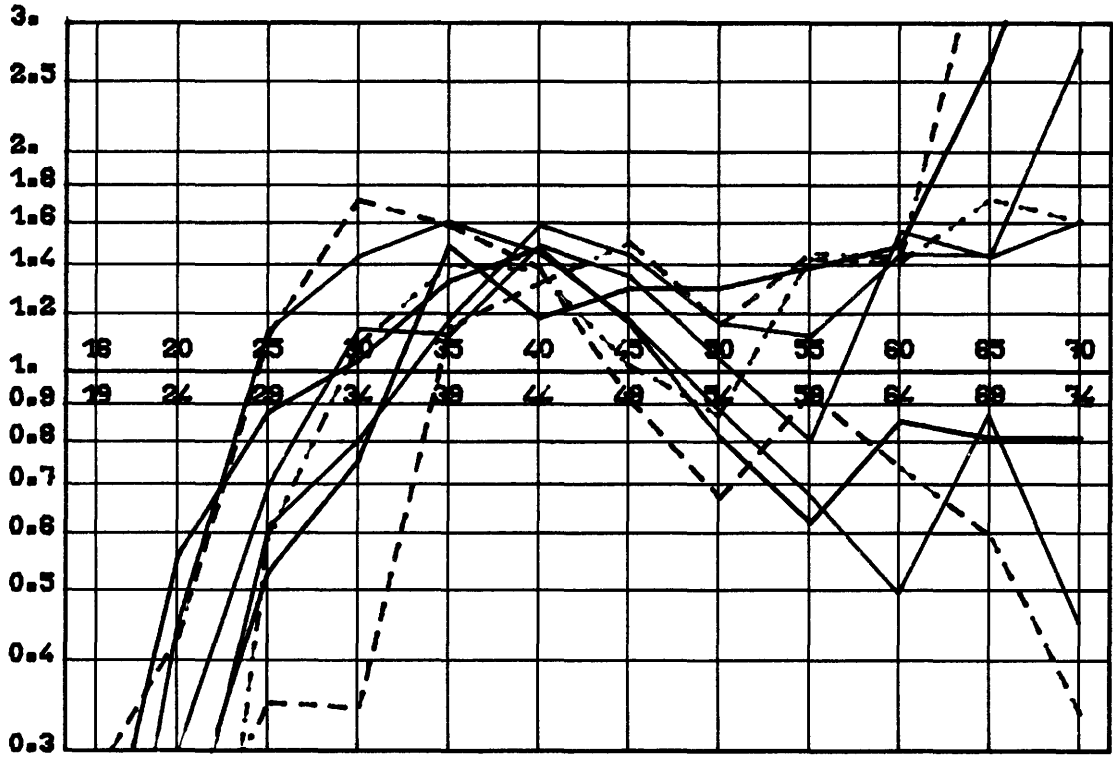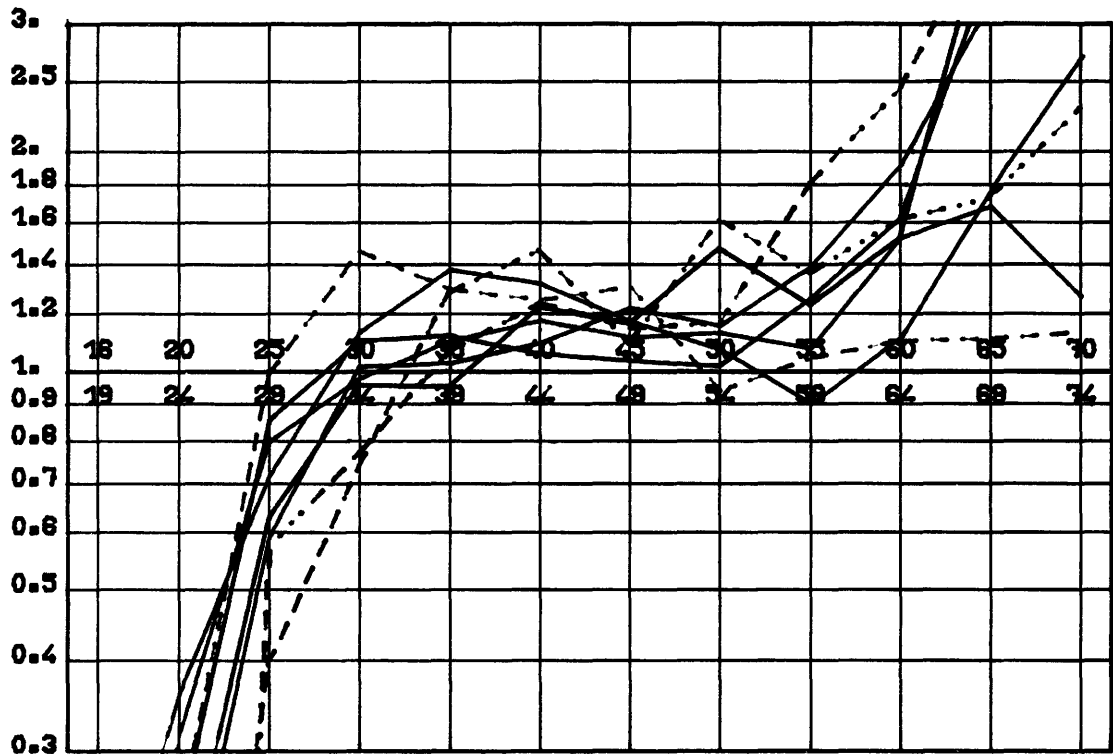
the oldest persons are much less pronounced ; for all countries the transition from under-representation to over-representation occurs around the age of 40. In the case of *employers and self-employed in the building industry* (see graph on previous page, above), one distinguishes two groups of countries : on the one hand Italy, Ireland, and the United Kingdom, which have under-representation of young people only under 25 (instead of 30 or 35 as in the other countries) and under-representation over the age of 50 ; there is also over-representation culminating around 30-40 years of age in the case of the latter two countries, and around 35-45 years of age in the case of Italy ; moreover, there is a group of six other countries where over-representation (coefficient of 1.2 to 1.4) just about covers the 35-65 age group, with under-representation below and more pronounced over-representation above. In what concerns *employers and the self employed in the distributive trades* (see graph on previous page, below) the profiles of the different countries are also rather similar ; in almost every country there is a level stretch between 35 and 60, under-representation being very pronounced only below 30 years of age ; in a less pronounced fashion than for agriculture, the United Kingdom is distinguished by a flatter profile. In the case of the other *employers and self-employed* (see graph above), the profiles of the various countries are fairly similar, with under-representation below the age of 30, a plateau, and then over-representation after the age of 50.

2. ANALYSIS OF THE DISTRIBUTION BY AGE OF THE EMPLOYED FEMALE POPULATION

2.1. The analysis of the distribution of the female population will be given in less detail than that of the male population and this for two reasons : on the one hand female employment is more concentrated in a limited number of groups, and in those groups having few workers the results are no longer significant, since the figures used come from a sample survey ; moreover, the differences between profiles are less pronounced than for males.

2.2. This last remark is translated immediately into a quantitative form if one examines the table giving mutual information for the countries, which are arranged here also in decreasing order :

Ireland	71
Luxemburg	60
Italy	44
Netherlands	25
Belgium	24
France	19
F.R. Germany	16
Denmark	16
United Kingdom	14

This mutual information is much slighter than that for men. For example, in the case of France and the Federal Republic, there are only one-sixth as many elements of information. The difference is indeed less pronounced (but still less than one-half) in Ireland, Luxemburg, Italy, and the Netherlands, but these countries have low female employment ; the sample is therefore smaller, which results, as we have already said, in an increase in mutual information.

2.3. If one now proceeds to the analysis of the main components, the contribution of the different axes is the following :

1st axis	50 %
2nd axis	18 %
3rd axis	10 %
4th axis	6 %

The results are therefore a little less favorable than are those for men, and this also because of the sample size.

In the following diagram, as in the case of men, the first axis opposes the young and the old, and the second middle and extreme ages, but in a slightly less distinct manner.

2.4. A more detailed examination of the graph leads to a classification that is not very different from the one we have selected for men ; first of all, the employees in distributive trades on the one hand, and those in banking and finance on the other hand
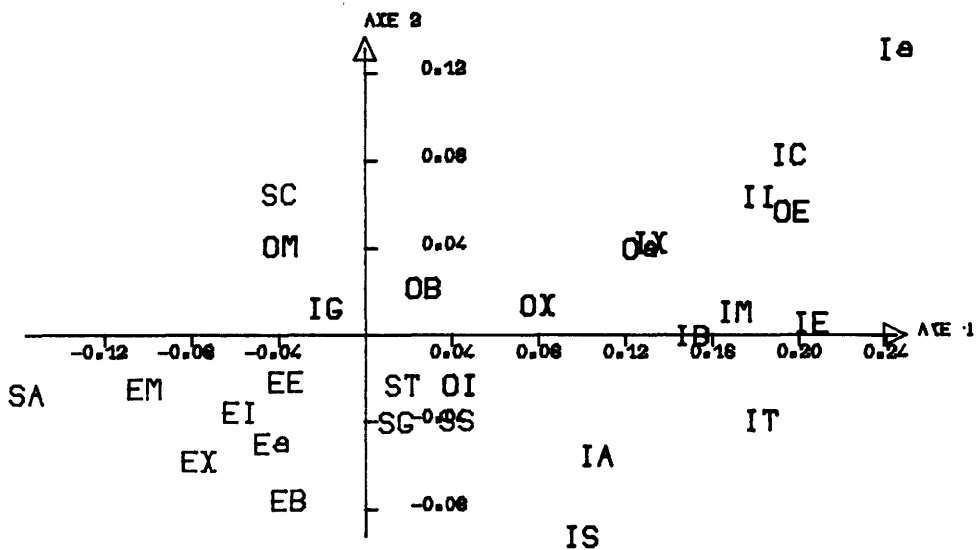
each constitute two relatively isolated groups ; next, among manual workers of industry, only those in the manufacturing industries should be isolated from the others, because of the greater relative importance of the very young (but we shall see that, even there, the difference is not great), the manual workers of energy and water being too few to merit being isolated ; the other employees of the services on the one hand, and the non-manual workers in industry on the other hand, each constitute a rather compact group ; finally, among employers and the self-employed, only those in distributive trades appear numerous enough to be analysed ; other employers and self-employed are therefore not treated nor are employees in agriculture.

2.5. *Employees in the distributive trades* (see graph below) have, with the exception of Italy, very similar profiles at least until about the age of 50, the over-representation of the very young 16-19 age group is situated between 1.5 and 2 and is therefore less great than that of men ; next, there is a small under-representation common to almost all the countries, between 25 and 40 ; the divergences that appear above the age of 50
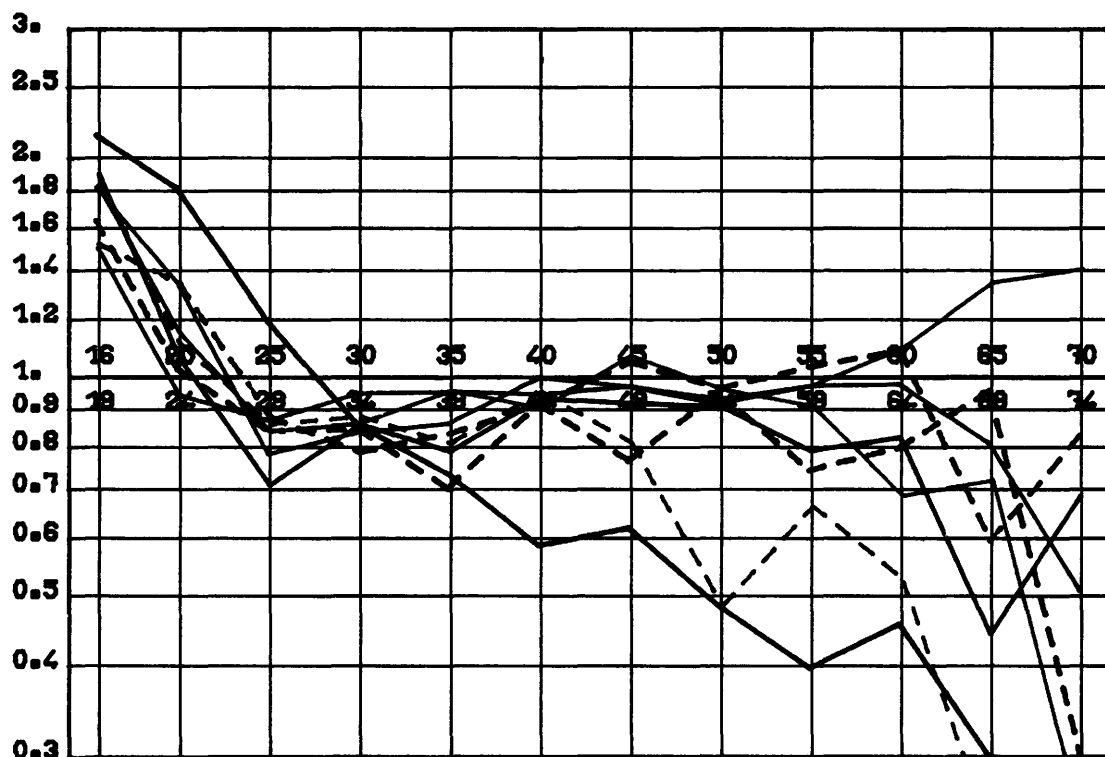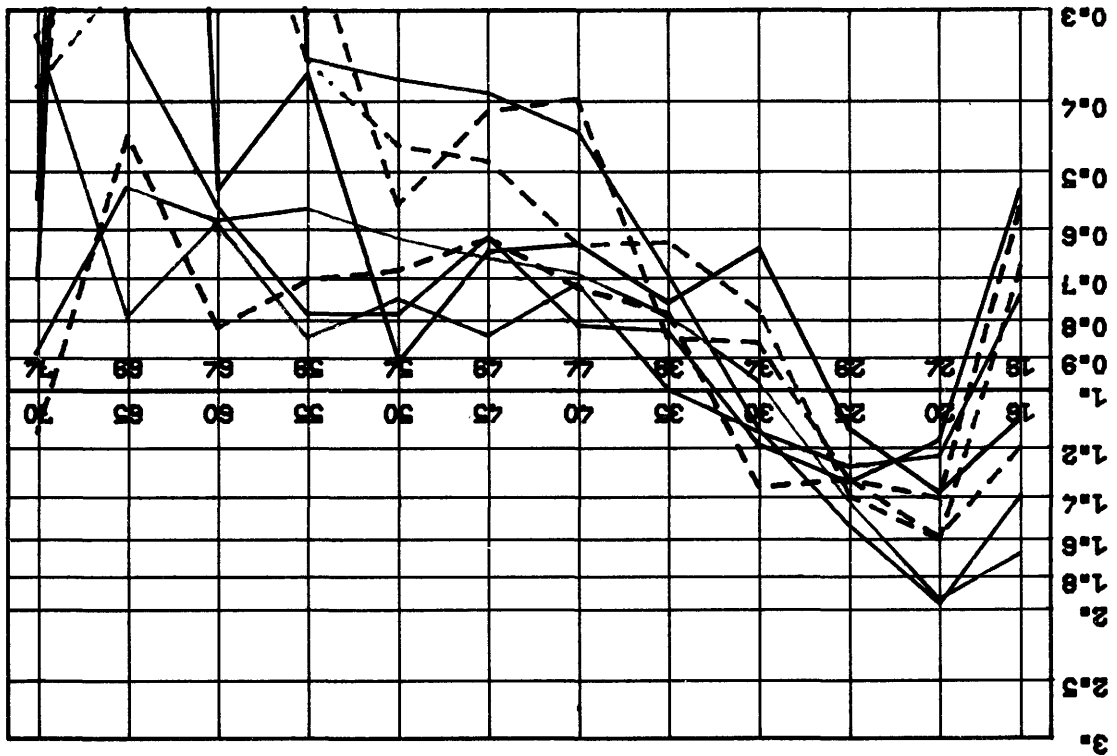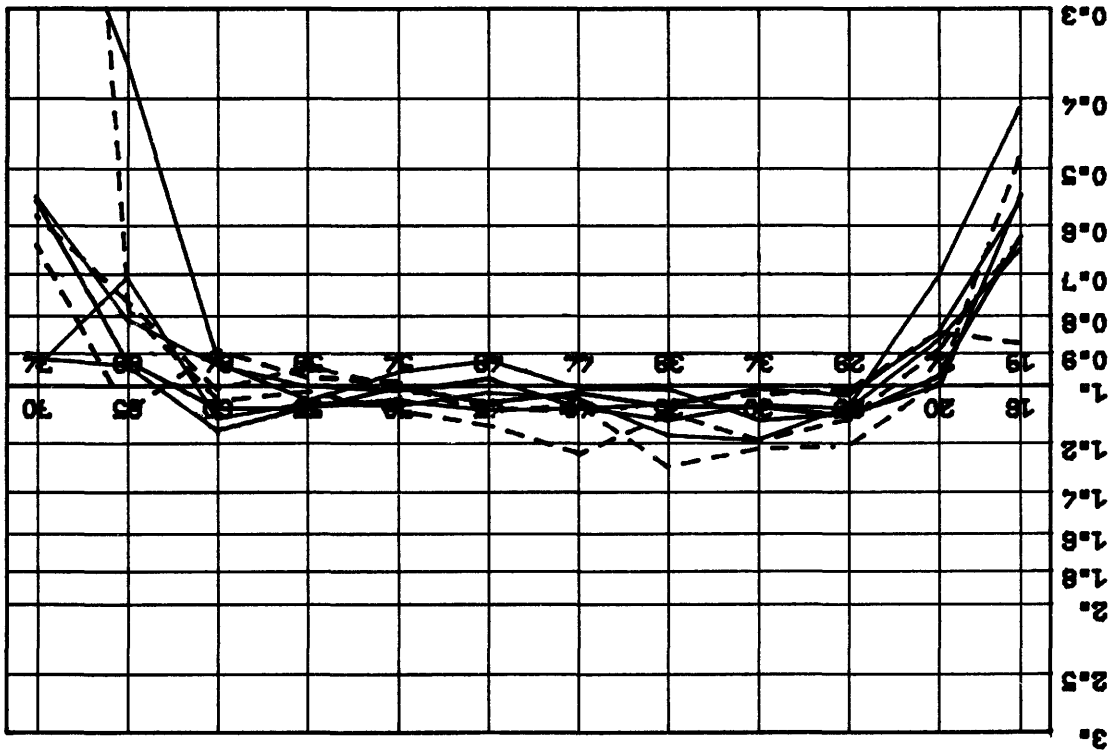
are not significant, except perhaps in what concerns over-representation in the United Kingdom and the very rapid decline in Luxemburg. Last, Italy's profile is very different from the others and very regular, over-representation beginning at a level above 2 in the case of the very young and declining regularly to under-representation towards the age of 30. The group of *employees in banking and finance* (see opposite graph, above) shows distinct differences from the analogous group of men, as one notes by comparing the graph above to the one on page 14. The under-representation of young people 16 to 19 remains true only for Belgium, Denmark, Ireland, and France, while there is over-representation in all the other countries ; in all the countries, nonetheless, the maximum of over-representation is situated in the 20-24 age groups or above. One should also note that from the age of 40 or 50 under-representation is very pronounced in Luxemburg, Ireland, and Italy. It is in the latter country that the contrast between male and female profiles is most pronounced. *Other employees in the services* (opposite graph, below) are for the most part employees of the different services which alone constitute 30 % of
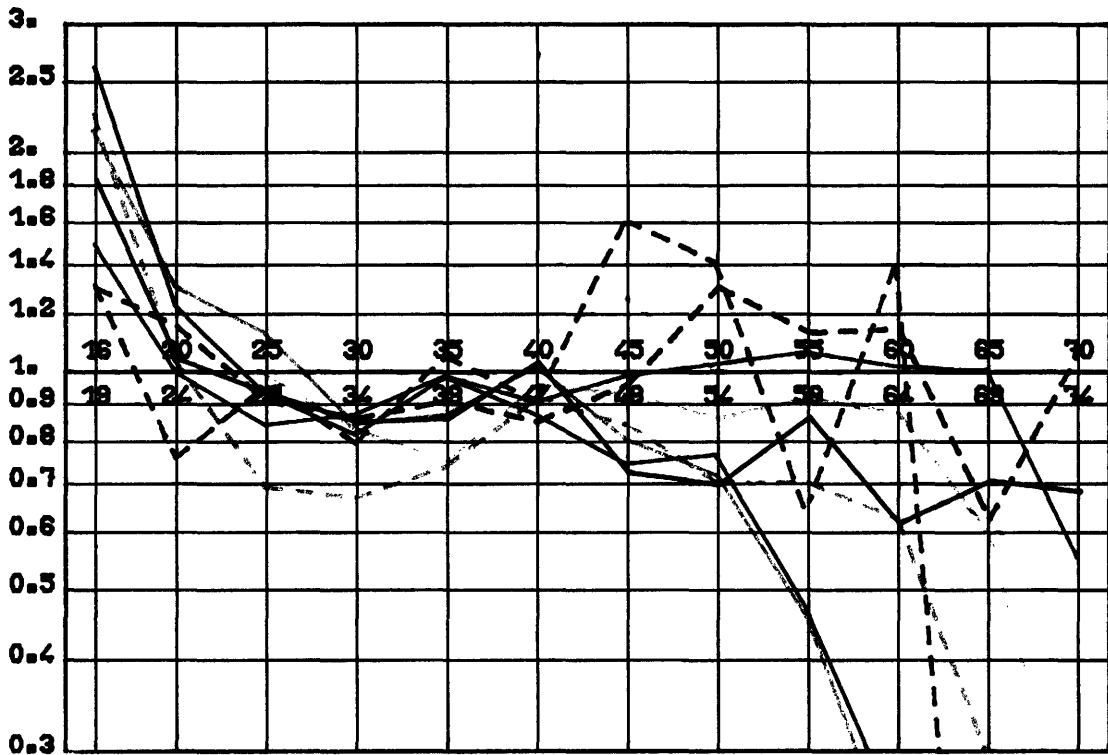
female employment, the entire group constituting 40 %. Given the technique chosen, i.e., the analysis of the relative profiles in relation to the country's total profile, a group having such an important relative weight can hardly differ from the whole ; it is probably rather heterogeneous, but there exists no information that would enable one to break it down. The 16-19 age group, contrary to what is the case in other countries, is not distinctly under-represented in the Federal Republic and Denmark.

2.6. In what concerns *manual workers of the manufacturing industry*, the profiles (see graph on previous page) are very similar to the corresponding male profiles but over-representation declines faster with age as it does in the preceding group. Over-representation in the first two age groups is particularly pronounced in Belgium, Italy, and France ; in the first two countries there is very great under-representation starting at the age of 55. In contrast, the profiles of Denmark, the United Kingdom, and the Federal Republic
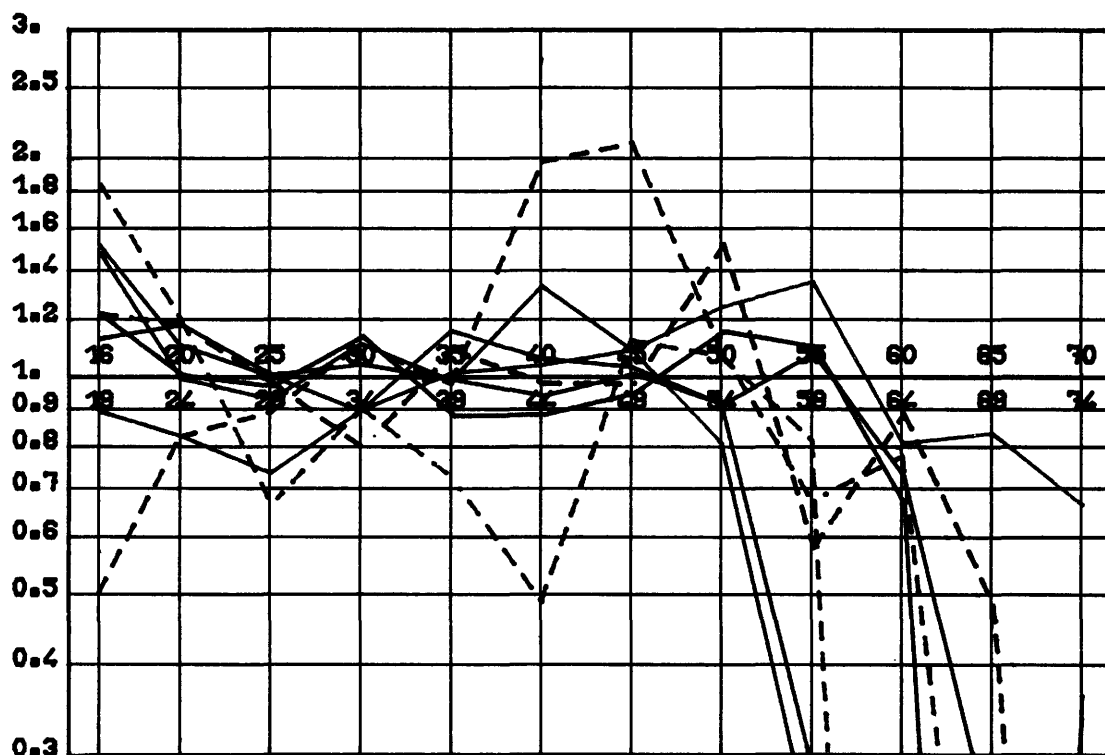
appear very flat. The other *manual workers in industry* (see graph opposite, below) conform to the average up to the age of 60 or 65, there being in most countries a slight under-representation of the very young. The exceptions to this typical profile are on the one hand Luxemburg and the Federal Republic, which have under-representation of young people under 25 and over-representation of the 30-55 age group, and on the other hand the United Kingdom, where there is under-representation below 35, then increasing over-representation culminating at the age of 60. Last, there is great over-representation of young people from 16 to 19 in Ireland, Belgium, and Italy. *Non-manual workers in industry* (see graph below) have rather diverse profiles. The cases of Ireland, Italy, and Luxemburg are characterized by distinct over-representation of the 20-30 age group and especially of the 20-24 age group in the two latter countries, compensated by a distinct under-representation from the age of 40 or 45. In the case of the other countries the differences are less clear and the trend can be approximated by a progressive passage from slight over-representation below the age of 30 to slight but more pronounced under-representation among the 16-24 age group in the United Kingdom.
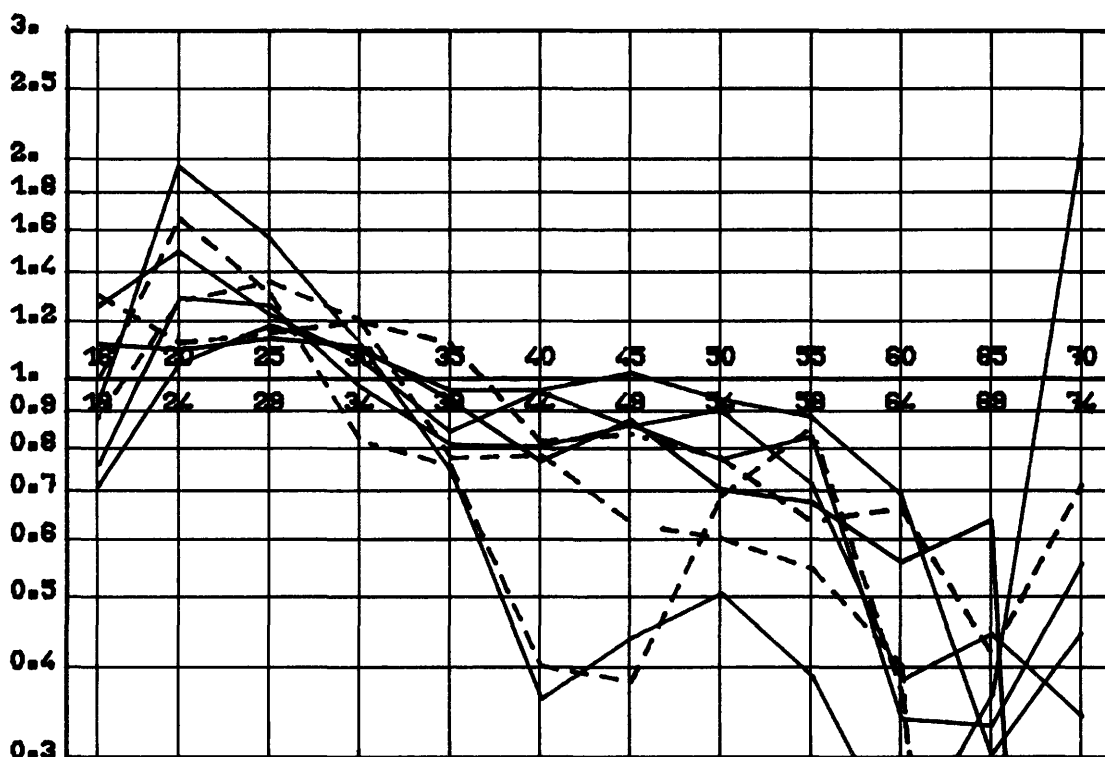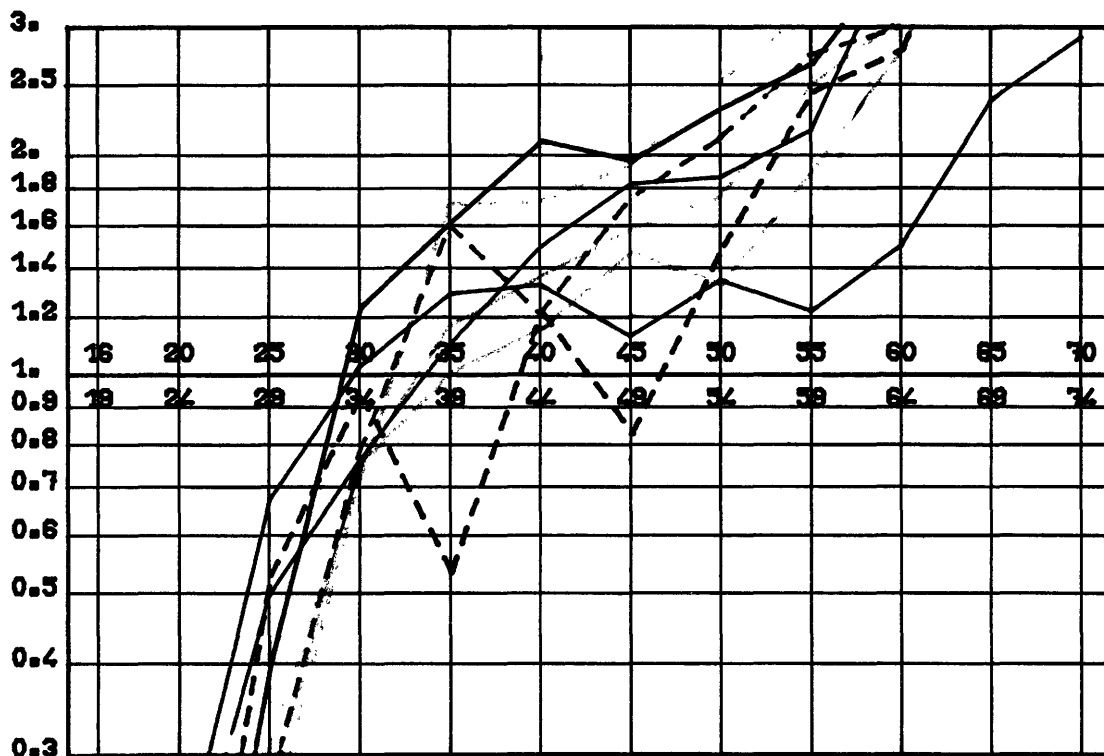
2.7. *Employers and the self-employed in the distributive trades* (see graph above) are largely under-represented below the age of 30, but nonetheless in a little less pronounced fashion than elsewhere ; inversely, over-representation is great at the most advanced ages, except in the United Kingdom ; the differences are rather pronounced between the Netherlands and Ireland where over-representation is already great over the age of 35, and France, the Federal Republic and Denmark where over-representation does not become important before the age of 55.

ANNEX : INFORMATION THEORY APPLIED TO ANALYSIS
OF CROSS-CLASSIFIED DATA

A.1. To choose one individual among x individuals, or to locate one individual among x classes — these classes making a partition —, one needs an information corresponding to the number of binary questions (yes-no) ; this quantity of information is therefore equal to :

$$log_2 x$$

This is the formula of Shannon ; the quantity of information is measured in bit (binary digit) ; by a change of scale, one may choose to measure it in neperian information unit, which makes some calculations easier, and therefore to take the measure of information :

$$ln\ x$$

A.2. Given x_0 individuals, divided into n classes making a partition, and considering that each class with the index j includes x_j individuals (wherefrom $\sum_j x_j = x_0$), the argument will be as follows :

If the distribution into classes in known, the necessary information will be :

— to locate the individual into a class j :

$$ln\ n$$

— to locate the definite individual into the class j :

$$ln\ x_j$$

or more exactly, as it is possible to result in any class j, the expectation :

$$\text{\&} \ (ln \ x_j)$$

wherefrom, in that hypothesis, an information equal to :

$$\text{\&} \ (ln \ x_j) + ln \ n$$

On the other hand, if the distribution into classes is not known, it will be possible to locate the individual only directly among the x_0 individuals, wherefrom an information equal to :

$$ln \ x_0$$

Knowing the distribution therefore brings out some information, which is the algebraic difference between both informations :

$$\text{\&}(ln \ x_j) + ln \ n - ln \ x_0 = \text{\&} \ [ln \ (nx_j/x_0) \ ] =$$

$$\sum_j \frac{x_j}{x_0} \ ln \ \frac{nx_j}{x_0} = \frac{1}{n} \sum_j \frac{x_j}{\bar{x}} \ ln \ \frac{x_j}{\bar{x}}$$

In the latter formula, $\bar{x}$ has the usual meaning :

$$\bar{x} = \frac{1}{n} \sum_j x_j = \frac{x_0}{n}$$

This quantity is called, depending on authors, the negentropy, or the deviance with regard to the equidistribution (there is equidistribution when $\forall j, \ x_j = \bar{x}$).

A.3. If the individuals are now cross-classified into two groups of classes (each making a partition), with the indices i and j respectively, one may write :

— the number of individuals belonging at the same time to the class i ($i = 1,2, ..., m$) in the first group, and to the class j in the second group ($j = 1,2, ..., n$) is :

$$x_{ij}$$

— the total number of individuals belonging to the class i is :

$$\sum_j x_{ij} = x_{io}$$

— the total number of individuals belonging to the class j is :

$$\sum_i x_{ij} = x_{oj}$$

— the total number of individuals is :

$$\sum_i \sum_j x_{ij} = x_{oo}$$

The negentropy corresponding to the cross-classification is :

$$\text{\&} \left(\ln \frac{m\, x_{ij}}{x_{oo}} \right)$$

The negentropy corresponding to the classification according to one of the groups of classes is :

$$\text{\&} \left(\ln \frac{m\, x_{io}}{x_{oo}} \right)$$

and to the other group, it is :

$$\text{\&} \left(\ln \frac{n\, x_{oj}}{x_{oo}} \right)$$

Knowing the cross-classification besides both classifications considered separately then corresponds to an information :

$$\text{\&} \left(\ln \frac{mn\, x_{ij}}{x_{oo}} \right) - \text{\&} \left(\ln \frac{m\, x_{io}}{x_{oo}} \right) - \text{\&} \left(\ln \frac{n\, x_{oj}}{x_{oo}} \right) =$$

$$\text{\&} \left(\ln \frac{x_{ij}\, x_{oo}}{x_{io}\, x_{oj}} \right) = \sum_i \sum_j \frac{x_{ij}}{x_{oo}} \ln \frac{x_{ij}\, x_{oo}}{x_{io}\, x_{oj}}$$

This quantity is called mutual information.

A.4. The definition of mutual information involves the taking into account of the structure coefficient :

$$a_{ij} = \frac{x_{ij}\, x_{oo}}{x_{io}\, x_{oj}}$$

which may be defined in several ways and has interesting properties.

If one considers that the classification is shown as a table composed with m rows i and n columns j, it is possible, writing :

$$a_{ij} = \frac{x_{ij}}{x_{io}} : \frac{x_{oj}}{x_{oo}}$$

to describe a_{ij} as the percentage of individuals belonging to the class j in the row i, divided by the percentage of individuals belonging to j among the whole population ; it is therefore a measure of the over- ($a > 1$) or under-representation ($a < 1$) of j into i.

It would also have been possible to write :

$$a_{ij} = \frac{x_{ij}}{x_{oj}} : \frac{x_{io}}{x_{oo}}$$

which indicates that a_{ij} measures the same way the over- or under-representation of i into j. One therefore has at one's disposal a coefficient which, unlike the very often used coefficients that are the row proportions x_{ij}/x_{io} or the column proportions x_{ij}/x_{oj}, can be used equally for comparisons of rows and comparisons of columns.

It is possible too, considering :

$$\overset{\frown}{x_{ij}} = \frac{x_{io}\, x_{oj}}{x_{oo}}$$

that is to say another distribution which would be strictly proportional to the distributions in both of group of classes, to write :

$$a_{ij} = \frac{x_{ij}}{\overset{\frown}{x_{ij}}}$$

which is the proportion between the actual number and the number that would exist in case of proportional distribution. It is then possible to notice that mutual information may be considered as a deviance of the real distribution with regard to a distribution proportional to marginal distributions, that is to say to the distribution in each group of classes considered separately.