A Correspondence Analysis of a XVth Century Census: the Florentine Catasto of 1427

Christiane Klapisch* **and Michel Demonet**
Ecole Pratique des Hautes Etudes, Paris

The quantitative historian inevitably finds himself attracted by vast bodies of material, but must often admit defeat before the sheer volume of such sources. Even when ordered in tables, the information so painfully collected often proves to be virtually unusable. The search for correlations among variables quickly overwhelms him with pages of coefficients which finally leave him with the inescapable conclusion that he cannot control his data sufficiently. In order to "unravel" it in its totality, to distinguish and classify subtle but important differences and, finally, to permit him to test his presuppositions or develop new working hypotheses, the historian need to have recourse to a more sophisticated statistical method — what we shall call here « correspondence factor analysis ». The purpose of this article is to show the effectiveness of this technique based upon the experience we have recently had in applying it to a medieval census containing economic, social and demographic data, the Florentine Catasto of 1427, one of the most difficult aspects of which has been the effects of geographical differences.

This research is part of a larger systematic investigation of the Catasto being made in collaboration with Professor David Herlihy,[1] one of the purposes of which is to study the function and structure of the family in Tuscan society at the beginning of the fifteenth century. The Catasto permits an exploration of the relationships between the familial and socio-economic

* I would like to express my gratitude to Emily Coleman, University of Pittsburgh, for the help she gave to me and for preparing this translation.

1 This project has been supported by the *National Science Foundation* in America and by the *Centre National de la Recherche Scientifique* and the *École Pratique des Hautes Études*, 6th Section, in France.

structures with a precision one can rarely find in western Europe before the beginning of the contemporary period.[2] It grew out of a reform of the fiscal system which the Florentines were forced to adopt in May 1427, during the war with Milan, when they found themselves pressed financially; and it has survived as a massive description of property and people. Its purpose was to inform the Florentine Commune so that it could distribute the burden of taxes more equitably and uniformly. The Florentines liked the idea that their fiscal charges did not depend solely on revenues from lands or moveable property and taxable labour, but that unproductive "mouths" might be used as tax deductions. Therefore, all of the approximately 60,000 household heads who were under the authority of the Commune had to declare precisely the nature and value (or rent) of their property, both moveable and immoveable (including commercial and private investments and debts, holdings in the public debt, merchandise and cash, animals and slaves, etc.). Moreover, they were required to provide a detailed description of all members of their households, indicating sex, age, relationship to the household head, state of health, and occupation. In all, 264,000 lay individuals were included in the survey.

The basic unit of the census was the taxable household, which the document consistently cited by residential parish. Furthermore, the *Catasto* is broken down into convenient sectors or subjects for study: individuals, families and households, parishes, townships, regions and the whole community of the Florentine state. Our research using correspondence analysis was originally based on the village or township; we then extended the analysis to the 146 rural cantons (*pievi* or *podesterie*, according to the administrative categories of the period). Our purpose was to clarify the different forms of rural households so that we could determine their responses to economic and social stimuli. Some of our studies to date, the most detailed being for Florence itself, had shown the links between wealth and the structure — more or less complex — of the households.[3] But did these relationships also exist in the rural society of Tuscany, and were they influenced by geographic differences in the countryside which encompasses plains, hills and mountains between the Tyrrhenian Sea and the Apennine ridge? A factor analysis, or what we call correspondence analysis, of these relationships allowed us to outline and interpret the regional groups that the data define for us, more precisely than would otherwise have been possible.

---

[2] For brief presentations of the data in the *Catasto*, see C. KLAPISCH, « Fiscalité et démographie en Toscane (1427-1430) », *Annales: Economies, Sociétés, Civilisations*, XXIV (1969), 1313-1337.

[3] See D. HERLIHY, « Vieillir au Quattrocento », *ibid.*, 1338-1352; IDEM, « Mapping households in medieval Italy », *The Catholic Historical Review*, LVIII (1972), 1-24; C. KLAPISCH, « Household and Family in Tuscany in 1427 », in *Household and Family in Past Time. Comparative studies in the size and structure of domestic group over time*, ed. P. Laslett, Cambridge, 1972, 267-281.

Our material came from a computerized edition of the *Catasto*; like the original document itself, it followed the inquest by households.[4] The data it contained were then aggregated for the population by each canton. The hundred or so variables computed in this way have been entered in a large contingency table showing, for each of the 146 rural districts, the number of times similar characteristics appeared.[5] Some of these data belong to groups of variables characterizing such topics as wealth, type of household structure, characteristics of its head, etc. How did the correspondence analysis handle this material?

Considered in its general aspects, factor analysis takes a matrix that is the representation of a cloud, or concentration of points, in a space of n dimensions, where n is the number of characteristics or, conversely, the number of subjects. It is difficult to illustrate this cloud; certainly one can construct as many diagrams as there are coupled subjects or variables, but the result will be useless. The purpose of descriptive statistics is to allow us to interpret a mass of data while losing as little as possible. In our case, the analysis was intended to show the cluster in a more limited set of dimensions, while still minimizing the loss of data. An axis which has the different points most dispersed along it carries the most information on the structure of the cloud: it will be the axis of least squares. The "second" axis will be the axis perpendicular to the first which best minimizes, in its turn, the loss of information; and so on for as many axes as are necessary.

Having determined its axes, "classic" factor analysis can make them "rotate". When the axes rotate together so that, resting orthogonally, each of them "explains" in the best possible fashion a cluster of intercorrelated variables in the cloud, this is called orthogonal rotation of simple structure; if each of the axes rotates separately so that it adjusts to one of the groups of variables, it is called simple structure oblique rotation. Before this rotation, the *factor* is one or the other of the axes determined by the least squares; after rotation, it is an axis adjusted to a cluster inside the cloud. How does one interpret these "factors"? Must one see them as underlying causes which can only be understood by identifying and naming them? For our purposes, we have used this method simply to describe the phenomena and to distinguish the relationships among the variables; the fact of naming the factors identifies no causality as such; it only allows us to relate our understanding of the phenomena.[6]

---

[4] This edition is available at the University of Wisconsin at Madison, and at the *Centre de Recherches Historiques de l'École Pratique des Hautes Études*.

[5] Only one series of information consists of means: the mean age of marriage in the canton, calculated for men and women following the method used by J. HAJNAL in « Age at marriage and proportions marrying », *Population studies*, 7 (1963), 111-136.

[6] For an introduction to factor analysis, see R. RUMMEL, « Understanding factor analysis », *Journal of conflict resolution*, 11 (1967), 445-480, and H. HARMAN, *Modern factor analysis*, Chicago, 1967.

The primary point of originality in the method of correspondence analysis is the use of the X^2 metric.[7] It is, in effect, theoretically possible to consider as interchangeable or reversible the list of subjects and their characteristics; thus, one can make two analyses of which one, the « Q-factor analysis » studies the variation of the subjects across their characteristics, and the other, the « R-factor analysis », defines the variation of the characteristics. The X^2 measure has the virtue that the factors are the same in both. In practice, only one analysis is necessary, and one deduces the other from it. For example, the totality of the subjects situated to the right of the first axis is determined by the characteristics also located to the right of their first axis. (Such a method obviously precludes all rotation of the axes in one or the other of the analyses, since this operation will destroy the correspondence between the factors of the two spaces.) The X^2 metric possesses still another advantage which is evident in the formula given in footnote 7: one can study the profiles of distribution by examining relative values rather than absolute values. It is not the uniformly distributed characteristics, even if the number of occurrences is important, which retain our interest, but the relative deviation of each of them among the subjects.[8]

In the same way that the percentage of total variance is calculated in "classical" factor analysis, we here calculate the percentage of total information given by the knowledge of the projection of the points in the cloud on one of the axes; we take this axis as equal to the moment of inertia in relation to the centre of gravity. In the same way that one calculates the coefficients

---

[7] This metric is defined by the following formula:

$$d^2(M_1, M_2) = \frac{\sum_j \left(\frac{n1j}{\sum_j n1j} = \frac{n2j}{\sum_j n2j} \right)^2}{\sum_i \sum_j n\,i\,j}$$

where M_1, M_2 would be two points, nij the number of times that the caracteristic j would appear in the subject M_1. The metric of X^2, a distributional metric, obliges us to consider the elements of our initial matrix as frequencies. Furthermore, the correspondence analysis does not use a matrix of correlations but a matrix of inertia. In reality, correspondence analysis is more a variant of principal component analysis than of factor analysis properly speaking.

[8] See particularly J. P. Benzecri et al., *Analyse des données*, Paris, Dunod, 1973, 2 vols (I: La taxinomie; II: L'analyse des correspondances). After being applied to linguistics, lexicology, psychology and pedagogy, to economics and to political history (cf. particularly A. Prost and C. Rosenzweig, « La chambre des députés, 1881-1885 », *Revue française de sciences politiques*, 21 (1971), 5-50), the method is beginning to be applied at the *Centre de Recherches Historiques* to historical censuses (military archives, the *Catasto*) and to various bodies of material (data on literacy in the XVIIIth and XIXth centuries; cf. F. Furet and W. Sachs, « La croissance de l'alphabétisation en France [XVIIIe-XIXe siècles] », *Annales, E.S.C.*, 29, 1974, p. 714-737).

of correlation between the different characteristics and the factors, we calculate the « relative contribution » of each characteristic to each of the axes; this carries the same kind of information on the strength of their relationship.

For this article, we will take an example of correspondence analysis, delineated by the framework of the cantons, expressly for the purposes of explanation. Actually we prepared the results by other means but within the same administrative outline, and we find them clarified by this type of analysis. Of the approximately 100 variables used, about half were demographic and the other half composed of social and economic indicators. The districts defined and classified themselves by their administrative characteristics, their geographic profile (site, altitude, distance from Florence or from the sea . . .) or by another, overall characteristic (area, size of the population, etc.). It is perhaps useful here to recall that our principal objective was to elucidate the variance between the familial structure and the wealth and social status in relation to environment. From the simple cross-tables and the computer mapping of the variables by canton,[9] we began to suspect the importance of wealth in the rural milieu and the importance of the type of land-holding as well. They emphasized the distinctiveness of the centre of Florentine territory, the *contado* (the lands longest under the control of the city) as opposed to the outlying "district", the regions most recently brought under submission. It is thus that we saw that the greatest proportion of "multiple" households, where several families live together, seemed to be among the richest classes and among the "share-croppers" (*mezzadri*). Particularly numerous in the *contado*, the *mezzadri* were, paradoxically, characterized by a total absence of capital, in terms of land: they occupied the property of landowners who were generally urban, and they brought with them only their labour. This anomalous alliance of penury and complex family structure merits some explanation. Correspondence analysis permits us to clarify to some extent this apparent contradiction by showing how these traits helped to mould the human landscape of fifteenth century Tuscany.[10]

Figure 1 shows in a simplified fashion the first plane which the two first factors define.[11] This alone summarizes a large part of the total data; it

---

9 The computer mapping has been done by the *Laboratoire de Cartographie* directed by Jacques Bertin at the E.P.H.E.; we particularly wish to thank Jacques Bertrand for his invaluable help.

10 These problems are developed in C. KLAPISCH and M. DEMONET, « A uno pane e uno vino. La famille toscane du début du XVe siècle », *Annales, E.S.C.*, XXVII (1972), 873-901.

11 To be more precise, we have eliminated from the two graphs the data which were most grouped at the centre and which were not discriminating in order to make the whole more readable. All together we eliminated about fifteen pieces of data.

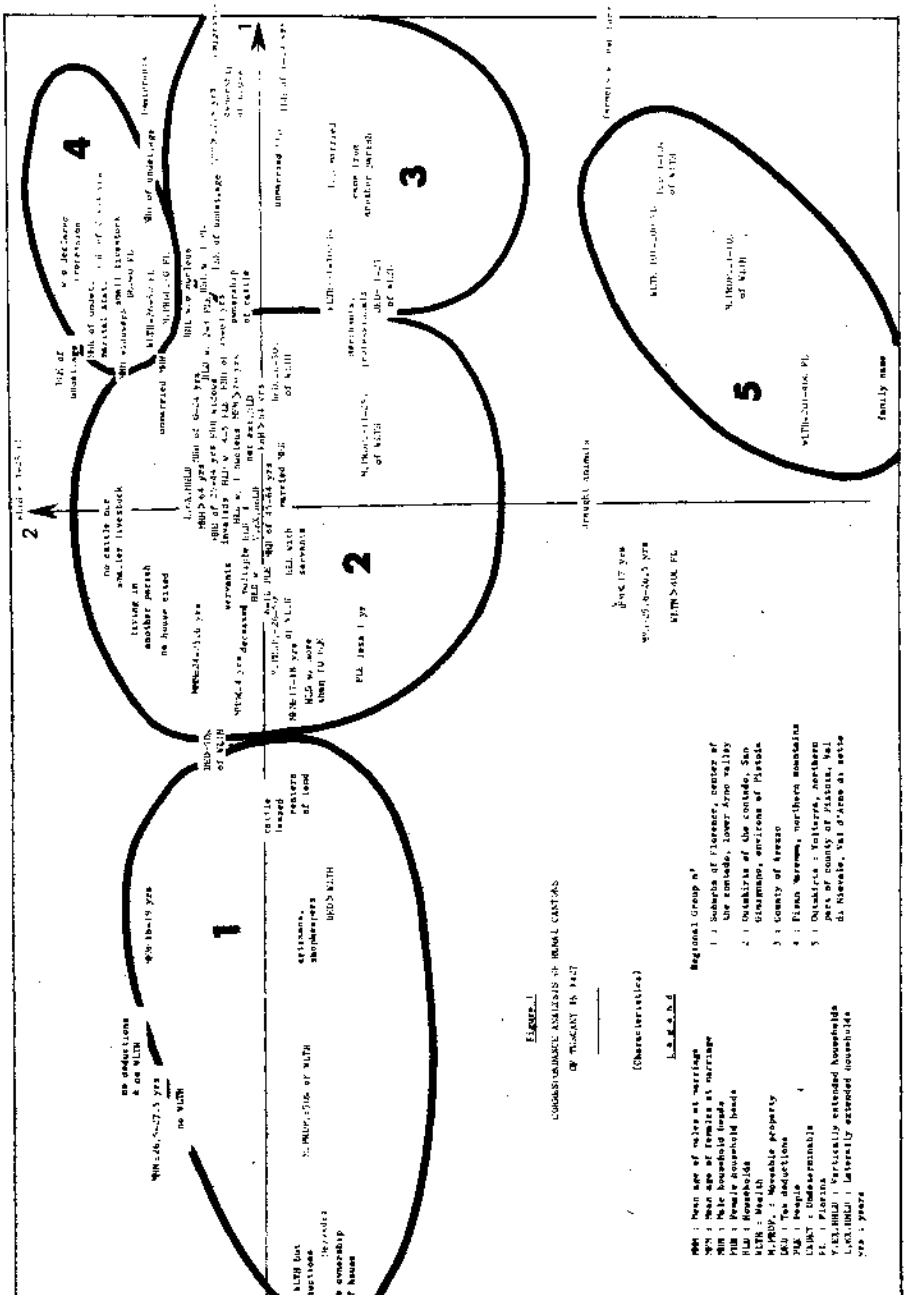

Figure 1. Correspondence analysis of rural cantons of Tuscany in 1427.

contains nearly half (48.7%). In order to understand better, we should again point out that the proximity between two « characteristic-points » in the figures indicates an association between the data they represent; between a « subject-point » and a « characteristic-point », their interaction; between two « subject-points », the resemblance between the behaviour "profiles" of the two subjects. In order to make the graph more readable, we represented the clusters of subject-points by numbered circles, the graphic equivalents of which are found in the legend on the graph.

The first two axes organize all our characteristics and subjects in relation to wealth. The first axis, the horizontal on the graph, contrasts in the aggregate those variables indicating the destitution and the indebtedness and absence of taxable capital to the left, with those that imply possession of goods to the right. This in itself accounts for 36.6% of the total information contained in the table of data. Connected to the two opposing groups of propertied and non-propertied individuals, occupational variables permit us to draw their social contours. To the extreme left are the *mezzadri*, who rent their houses from landlords, are burdened with debts and possess no other capital than their own hands. To the extreme right are the peasants who farm their own lands. Closer to the centre of gravity where the axes cross, one finds the farmers who rent their lands, the rural artisans and shop-keepers, all drawn to the side of the deprived; to the right, one finds the merchants and those "professionals" such as notaries, lawyers, physicians, who received a richer patrimony. Other variables are correlated to these groups: deductions equalling, or even surpassing the raw fortunes, characterize, at the bottom of the social scale, the scorned *lavoratori* of the countryside. And the renting of larger livestock is clearly linked to the renting of lands, while the gauges of household mobility are concentrated for the most part on the side of the small landowners.

The second axis, which describes only 12.1% of the total information goes from the greatest wealth, at the bottom of the diagram, to the lowest taxable category of people, possessing a few florins, at the top. An isolated characteristic, that of the presence of a family name, constitutes the inferior pole of the axis and we find it linked to wealth greater than 100 florins. As a consequence, the axis distinguishes the group of landholders. At the bottom, that is to say on the side of the wealthier, it clusters the peasants with their own lands and attracts in this direction the small group of merchants and members of « service professions ». Finally, at the top right-hand side, it underlines the relationship between the variables which indicate the small size of the property, the presence of isolated men and women at the head of a household, the small scale of the familial group and the frequency of migratory exchanges with the exterior world.

The first plane thus highlights, with the economic variables, the profound division of the rural society between landowners and non-landowners and

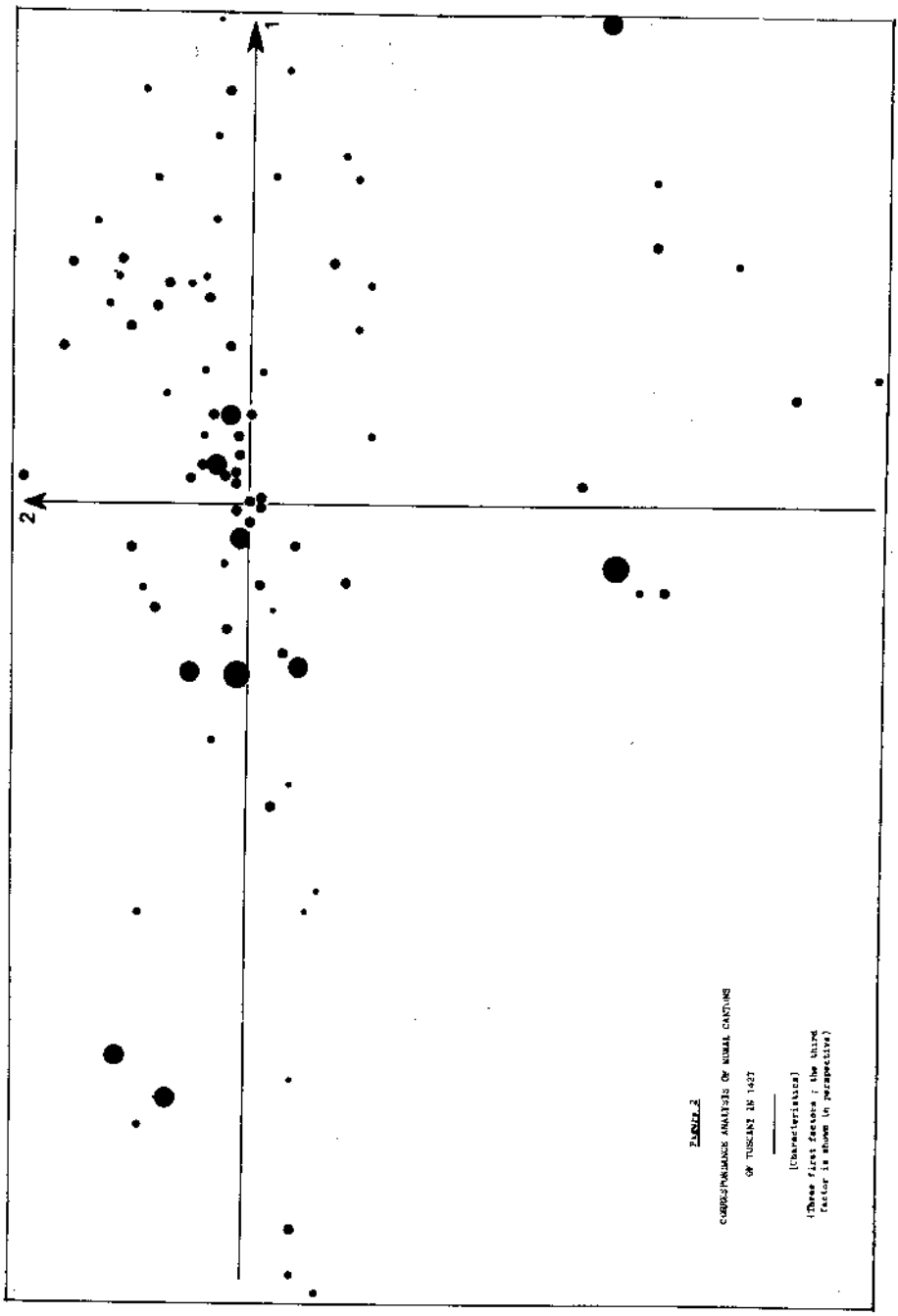

Figure 2. Correspondence analysis of rural cantons of Tuscany in 1427.

the secondary importance of the hierarchy of wealth. Do the variables describe the size and structure of the households as well? First one notices that they group themselves rather strongly around the centre. The diagram indicates, however, that while a certain relationship between the size of the family group and its wealth exists, this relationship is not linear. The characteristics which describe the largest and most complex households sit between the pole of the deprived, to the left, and that of the wealthiest, at the bottom. The smallest "*ménages*" find themselves lodged in the upper righthand quarter of the graph, associated with moderate wealth. Thus a certain polarization marks the familial structures since the largest domestic groups are found at both extremes of the rural society: among those who have nothing in the eyes of the communal fiscality and among those with the greatest taxable wealth.

An examination of the third axis allows for a more precise association betwen economic and familial characteristics. We have suggested, by a representation in perspective, the position of the variables along this axis; this is in Figure 2 which describes the first factorial plane only in terms of characteristics. The reader will be able to imagine then that the largest points, the strongest on the third axis, are the closest to him in a three dimensional space, while the smallest and the most "negative" on the axis would be the furthest from him, on the other side of the plane created by axes 1 and 2.

The third axis underlines the relationship between a certain number of variables describing the household and those which denote the mean ages at marriage for the men and women of the canton. It runs from the latest marriage ages, associated with the most reduced households and the professions least tied to the land, to the opposite pole which is defined by the youthful marriages of men and women and large household size. These characteristics are indicated on the graph by the thickness of the point. The group of peasant landowners is clearly attracted to this size; on the other hand, the *mezzadri* are associated with a household size and marriage ages that are less eccentric.

However, the weak autonomy of these familial characteristics is shown by the slender amount of the total data the axis expresses (7.3%). Furthermore, if the household size is obviously correlated with the age of male marriage, it is this last variable which determines the third axis; the indicators of the size and structure of the household group remain relatively clustered around the centre.

This analysis of the "characteristics" suggests that altogether the rural districts of the Florentine state were differentiated mainly by the relative presence or absence of capital. The primary factor in the dissimilarity of different areas of the countryside derived from the role that urban capital, the promoter of the *mezzadria* system, played; where the lands were in the hands of bourgeois townsmen, the rural classes appear to have been destitute of all taxable goods and heavily indebted to the landlords and moneylenders.

But the largest households are not uniquely associated with the wealthiest milieu. The household size characterizing the whole of the group of *mezzadri* clearly passed the mean. The larger patrimony certainly encouraged the family group to grow and diversify; but the assurance of cultivating property sufficient to support a family, even under the menace of a short-term *mezzadria* contract, appears to have also been an important stimulant.

Figure 1 represents, in a summary fashion, the main clusters of subject-cantons, superimposing them on the characteristics of the first and second axes with which they are associated. In this way, one can distinguish, to the left and towards the centre, the Florentine *contado* as opposed to the quasi-totality of the "district", which one finds almost entirely to the right. The cantons which are near the characteristics indicating the absence of taxable capital, the presence of the *mezzadria* and indebtedness, localize around the periphery of Florence, on the plain between Prato and Pistoia, and especially to the south of the capital and of a line going NE to SW (Florence-Castelfiorentino) which seems to be the furthest limit of the extension of the *mezzadria* system. The rest of the *contado*, the cantons either on the outskirts, which were still part of the hill or lower mountains zone, form a group well centred around the axes. The right of the graph brings together almost all the district. The north of the county of Arezzo, the region of Cortona and Castiglione Fiorentino, the small harbours of the Pisan county (Livorno, Castiglione della Pescaia) all are situated to the right of the first axis, in a zone characterized by moderate wealth, an abundance of migrants, and small or truncated households. All these regions are, in effect, on border areas of Florentine territory or else in the mountainous areas where the small landowning peasants were able to maintain themselves — but which the war with Milan would have recently ravaged.

The second axis, which differentiates wealth, permits us to distinguish between the poor and the rich in the outlying regions of the Florentine *contado* and particularly in the district. Corresponding to the former, at the top of the graph, are the Pisan hills and a part of the east of the district. At the other extreme, we find the wealthiest, essentially in the NW of the Florentine territory: the Pistoian region, Val di Nievole around Pescia, Val d'Arno di sotto around Fucecchio, Garfagnana around Barga to the north of Lucca; the region of Volterra and the immediate environs of Arezzo follow. The appearance of family names, earlier in Val di Nievole than elsewhere, makes the region highly conspicuous, and it is moreover richer and more urbanized than the rest of the Florentine countryside. As a consequence, the second axis allots the highest level on the scale of wealth to all the regions not characterized by the *mezzadria* system, underlining the economic prosperity of the north-west.

The analysis allows us to distinguish among the main household types which characterize these regional groups. Large familial units and an early

age of marriage mark the area around Pisa and, in the radius of a large circle around Florence, the cantons of its periphery. The Florentine suburban areas, taking the term in a broad sense, the Pisan Maremma, the north of the county of Arezzo, on the other hand, place themselves on the side of smaller households and a later masculine age at marriage; these traits accompany high proportions of artisans and people involved in "services" in the midst of the rural population, and more frequent migration.

Figure 3 juxtaposes by columns the maps representing some of the essential characteristics of our factors and the co-ordinates of the cantons on each of the three axes. In this way, one can see the quasi-identity between the co-ordinates on the first axis and the presence of taxable capital which is itself correlated inversely with the proportion of the *mezzadri*. To the west, the *podesteria* of S. Maria a Trebbio, near the city of Pisa, alone escapes a parallel between the spread of the *mezzadria* system and the number of taxable units: the *Catasto*, in effect, does not categorize these wretched peasants as *mezzadri*. The second column presents some components of the second axis: modest patrimonies, the proportion of households benefiting from fiscal deductions, the appearance of family names. The third column selected certain characteristics which made up the third axis. This column underlines the inverse relationship between advanced male age at marriage and the presence of "multiple" households. But we see that, at the heart of the territory and particularly in the Arno valley downstream from Florence, the proportional increase in households does not keep pace with the clustering of several familial nuclei under the same roof. Did the possibility of remuneration for labour — either by salary or as a "share-cropper" — with a ready marketplace in the neighbouring large city, permit the fertility level to maintain itself and the conjugal households to attain a moderately elevated size, while the communal life of several families inflated the households of the less favoured regions? This is one of the new problems whose solution we shall have to seek by other means.

The correspondence analysis, some of the results of which we have schematically given here, has thus confirmed and enriched our working hypotheses. From the point of view of social and demographic history, it has corroborated — at the level of the canton — our impression that there was a non-linear relationship between wealth and the structure of the familial group; but even more it has revealed the central role that the *mezzadria* system played in the reconstruction of the family group at the end of the Middl e Ages. The rural familial structure certainly depended on the basic wealth of the household which, according to our document, was considered in terms of capital; a larger amount of property could evidently support a larger domestic group, hasten the marriage of the children, tighten the ties among family members. But the spread of rural wealth was too limited for this factor to be as important as it was an urban environment. On the

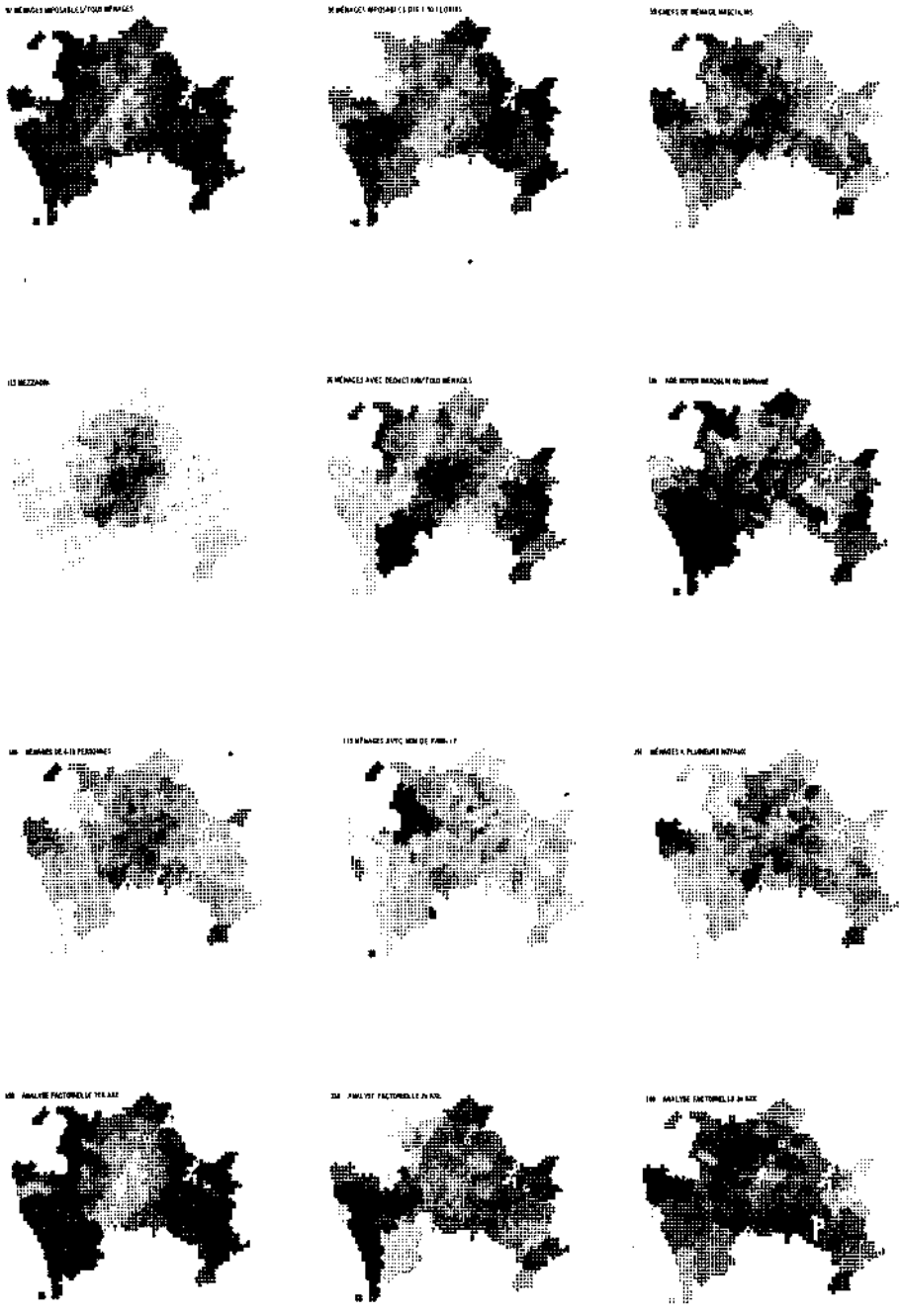

Figure 3. Computer mapping of some characteristics and the three first factors of the correspondence analysis.

other hand, at the beginning of the fifteenth century, the social status of the household and, finally, the landed property system and the penetration of urban capital into the countryside surrounding Florence played a primary role in stabilizing the rural family, which was reduced and broken by demographic catastrophes. They clearly encouraged the renewal of higher fertility rates, spotlighted for us by the youth of the population in the region of the *mezzadria* system, and the maintenance of intergenerational ties, marked by the frequency with which we find vertically extended multiple households. On the contraty, however, in those areas of small peasant landowners or where the land was held under an extremely long lease, emigration — a response to a crumbling peasant patrimony — rapidly destroyed these ties by forcing the departure of "superfluous mouths".

By this method, we have also been able to better encompass all the variables linked together and whose interconnections allow us to describe whole regions. In this respect, the political unities drawn by history occasionally prove very coherent, as in the cases of the Florentine *contado* or the county of Arezzo, and are occasionally singularly heterogeneous, like the old county of Pistoia. Thus new avenues into the problems of regional differentiation are laid open.

From the point of view of the methodology, some of the advantages most often stressed in factor analysis, in one or the other of its variations, is the ability to sort out the variables and spotlight their interconnections. We will take as an example our various indicators of wealth: henceforth, it is possible to retain only crude wealth as an index of fortune since the taxable wealth is closely linked to it at the level of the canton.[12] On the other hand, we can take the proportion of the deductions allowed by the Commune from the crude wealth and show that they are not linearly correlated; the reductions depend on the type of activity of the household head and the zone of residence; these reductions constitute a different indicator of the social level. The analysis permitted us to classify this index of economic dependence for which the hierarchy reappears as a principal component of the fourth axis.

We might be criticized for reuniting in a disordered fashion too many variables in relation to the number of subjects, and it might be claimed that they regroup themselves because they are closely interconnected and in reality measure the same thing. In order to counter to this objection, we later reduced the number of characteristic-variables and kept only a certain number

---

12 We would note that classic factor analysis which adjust its factors to the clusters on the inside of the cloud by rotation must avoid series of data being redundant since they would create clusters where axes would unfailing pass because the data themselves are not independent.

of indicators of wealth, socio-economic status or demographic and familial structures, which are not closely linked together. Preserving from the total set only a third of the initial data (33 variables) and avoiding the introduction of complete series of variables (for example the whole gamut of the categories of wealth), we arrived at the same configurations: variables and subjects cluster in almost exactly the same way as in our graphs 1 and 2, and the axes of inertia for this reduced range of characteristics and for the corresponding range of the subjects ordered themselves in the same manner as in the analysis using a matrix of all variables. Thus, we have made certain it holds a stable structure, at least from the point of view of the distributional metric.

These results show the flexibility of the method. "Classic" factor analysis does not generally have a good reputation among historians. It is accused of pushing its willing practitioners along the path of least resistance. Expecting that some ideas will come out of the mixture of their data once it has been shaken and churned within the computer, they content themselves with doubtful results, after having manipulated the "information" without question . . . Does correspondence analysis, which allows the description of a double body of data, in the most compressed and pertinent form, escape these criticisms? At the outset it must be stated that its results will remain largely indecipherable if there has been no preliminary study to pinpoint the problem that they are to unravel; the amount of intuition and empiricism that their interpretation requires are employed in vain without some working hypotheses. But these remarks do not concern only those who use this particular statistical technique. A historian must develop his questions as a function of what the data proposes to him and also by the selection and presentation he makes from the material. The successive refinements to which we have submitted the original corpus show, from this point of view, that correspondence analysis recommends itself as much for the preliminary treatment and the sorting of the whole mass of still poorly explored information as for the study of specific problems raised by the interpretation of the first results. Searching for a classification of the totality of the data rather than the disclosure of hidden causality, the method orders the information in a pertinent way and groups the variables into interpretable structures. It permits us to set forth some hypotheses which correct presuppositions and even to extract conclusions. Used at an intermediate stage, when the researcher wants to locate the groupings and eccentricities of his "subjects", the independence or interrelations of his "characteristics", correspondence analysis, a descriptive method, can also suggest some hypotheses that later statistical tests can verify.

---

[13] Cf. R. M. Zemsky, « Numbers and History. The dilemma of measurement », *Computers and the Humanities*, 4 (1969-70), 31-40.