Towards a geographical information system for European economic community regional data

Lane, Amanda M. J.

TOWARDS A GEOGRAPHICAL INFORMATION SYSTEM FOR EUROPEAN ECONOMIC COMMUNITY REGIONAL DATA

Amanda M.J. Lane

MSc. Dissertation, 1981

CONTENTS

ABSTRACT

The study proposes the development of a geographical information system for European Economic Community (EEC) regional data. The features of geographical information systems and some related issues are discussed. An outline is given of likely user requirements and attempts by various European organisations to make provision for these.

The practical work involves the application of three selected data analysis, manipulation and mapping packages, and operating system facilities to tasks which users of the data might wish to undertake. For this purpose, EEC "Level II" region boundaries were digitised, and selected socio-economic attributes for 1977 organised in disc files. Ten choropleth maps of the Level II regions (eight original variabels and two composite variables) are displayed. Although there is a lack of certain, more specialised features in the software system used, it is considered that there exists the imprtant basic requirements of data analysis, manipulation and mapping in package form, upon which to build the proposed geographical information system.

NOTES

Practical work was undertaken while there were only nine EEC member states. The admission of Greece as a tenth member state has been deliberately disregarded for the purposes of this dissertation.

The mapping package used was GIMMS, version 3. A number of suggestions for improvements are made in section 4.1, many of which have been implemented in GIMMS version 4, recently available. Section 4.1 also suggests a pre-processor interactive interface for CLUSTAN. I now understand that such a facility exists at York University, called "CLUSCOM".

INTRODUCTION

The EEC central data-gathering agency, EUROSTAT, currently provides socio-economic spatial data on a yearly basis, in book form, for regions of the European Economic Community (EEC). A regionalisation hierarchy has been developed which attempts to allow regional comparisons to be made over the nine member states. Given that research into large data sets is, in general, most efficiently carried out by computer, it would be of considerable advantage if a data base were to be set up, and data provision made to interested parties via some computer-compatible medium such as magnetic tape.

However, the existance of an EEC data base alone is largely useless unless some efficient means of accessing, manipulating and mapping the data is also available. the ideal goal may be considered to be a "geographical information system" for EEC regional socio-economic data, whereby researchers, decision-makers and administrators can access and manipulate whatever portion of the data base they require by "conversing" with a user interface program.

CHAPTER 1

INTRODUCTION TO GEOGRAPHICAL INFORMATION SYSTEMS & RELATED ISSUES

1.1 DATA MANAGEMENT BY COMPUTER - INTRODUCTION AND SOME TERMINOLOGY

There is generally little doubt of the utility of a computer system as a medium for data storage, analysis and retrieval. It has clear advantages over manual systems in that the data may be kept in compact form on magnetic disc or tape and this may be accessible at remote sites by a number of users working simultaneously, information retrieval, information up-dating and data analysis can be rapid and non-tedious, and maps and graphs may be generated automatically more or less as the user requires them. Having said all this, it must be emphasised that the task of setting up such computer systems tends to be time-consuming, non-trivial and may be costly, depending of course on the extent of the computing facilities already available. It is also important to bear in mind the differences in the errors which may occur between a manual card index retrieval system and a computer-implemented system. The former is more subject to random human error such as mis-reading names, the latter more subject to systematic error such as that caused by program bugs.

For these reasons, it is essential to evaluate clearly the scope of the requirements, because it may well not be worthwhile to implement such a system for a "once-off" project, such as the production of a single detailed map, or the determination of a single statistic. The benefit derived from computer management of information is probably at its greatest in an environment where the requirements include multi-user access from remote sites, the analysis of various portions of the data in a variety of ways, the graphical expression of the data in a variety of ways, and the reformatting, cross-referencing and tabulation of data on a large scale.

The term "computer systems" is used here in a general sense to include both the hardware and software necessary for data management. The main emphasis in this study, however, is upon the software aspect, which may be referred to as a system in itself. A number of terms are now in general use, all of which refer to some type of data organisation or management facility within the computer, and are used interchangeably by different people. These include "data set", "data base", "data base management system" and "information system"; unfortunately the distinction between them is far from clear. The following paragraphs attempt to provide some rough guidelines:

(a)Data Set. This is often used to refer to a single file of data, perhaps structured as a cases-by-variables matrix.

(b)Data Base. This usually refers to a number of interrelated data sets (as defined above) but may extend to include management programs written for data retrieval.

(c) Data Base Mangement System. This is a more specifically used term which refers only to the management software, as distinct from the data. Such management systems may vary in sophistication from one where the user is required to supply some of his own program routines to one where the user interface is a query program to which simple replies are all that is needed for data retrieval. These systems are usually general-purpose in nature.

(d) Information System. This term is very often used interchangeably with the previous one, but is increasingly used to imply a system of greater sophistication in that the user interface may be an easily understood process of query and reply and the system may provide not only data retrieval but data analysis, graphical display, data validation and other data manipulation processes. The term may sometimes be used to include the data as well, particularly when these systems are designed for a special purpose, built around a certain kind of data base.

The last of these loose definitions is the concern in this study. An "information system" shall also be taken here to mean only the software and not the computer and peripherals on which it depends, despite the temptation to include the latter where small, dedicated computing facilities are concerned. Although the creation and implementation of an information system may not be quick and simple, it is important that the use of the system has these qualities. The advantages of computer methods outlined in the first paragraph are based on a theoretical ideal, the practical reality of which depends greatly on the design of the information system, the efficiency of any software on which it may rely and the hardware supporting it.

1.2 GEOGRAPHICAL INFORMATION SYSTEMS

The type of data management facilities relevant to this study may be called "geographical information systems". The characteristics of geographic data are sufficiently exceptional to warrant a special case of information system. The following sections of this chapter aim to explain the facilities required of such a system, and to briefly discuss methods of meeting some of these requirements in terms of system design, software support and data encoding techniques.

1.2.1 Introduction And Outline Of Important Desiderata

Geographic data are particularly distinctive because they incorporate an important spatial element. The individuals to which the various attributes pertain should be considered in association with one-another and not in isolation, since they are spatially linked. An essential requirement for an information system designed to cope with such data is, therefore, that it should have available spatial operators and, in particular, should be capable of mapping, so that the variations in the data may be seen in true perspective. These geographical information systems in fact largely developed from the field of cartography as facilities for automated map production became available. The Ordnance Survey have transformed some 17000 of their maps into computer-compatible form, building up a digital topographic data base for automatic plotting. They are currently undertaking a project to re-organise the data base so that it may be linked with LAMIS (Local Authority Management Information System) to provide local authorities with important up-to-date topographic information relevant to the spatial data LAMIS already holds, for example on individual households. At present, local authorities have to cope by using ordinary paper map sheets sent to them by the Ordnance Survey and then transferring their data manually onto these sheets. By linking LAMIS directly with the O.S. topographic data base, local authorities could produce their own up-to-date O.S. maps with planning application positions or development plan zones already overlain.(See Thompson, 1979 and Harrison, 1979)

The current emphasis upon exploratory data analysis (Tukey, 1977) has particular relevance for geographers (Cox, 1978) and for other social scientists. Researchers are encouraged to explore their data, allowing the data themselves to reveal the underlying relationships, rather than take the classical approach of hypothesis formulation and subsequent testing. This exploratory approach is far more suited to the analysis of spatial socio-economic data, for example regional variations in living conditions, where the researcher may have few pre-conceived theories on which to base a hypothesis. Attempts to comform to rigid inferential analysis techniques by testing groundless hypotheses can be unhelpful and time-wasting when hypotheses fail to be disproved! It is far more efficient to allow the data themselves to reveal where the variations lie. The researcher concerned with spatial data must, therefore, have facilities with which to do some exploring, for example by being able produce maps at various levels of spatial aggregation. Geographical information systems can provide such facilities, enabling the user to perform quickly and easily the various different operations necessary to reveal any underlying relationships which may exist.

The precise requirements of a geographical information system will of course vary according to the user, but some fundamental capabilities for a good system may be outlined as follows (from Rhind, 1976):

(1) An efficient file structure, enabling easy cross-referencing and data matching to be carried out.

(2) Editing facilities for error correction or update both of attribute files and coordinate data files.

(3) Statistical analysis facilities including production of graphs and histograms, for example by linking to a statistical package.

(4) An ability to generate data output in tabular form in specified formats.

(5) An ability to retrieve information for any geographic individual

specified by name or coordinates, on any combination of attributes.

(6) An ability to find or to calculate descriptors from the coordinate data such as maximum and minimum coordinates of a string or area of a polygon.

(7) Facilities to build polygons from supplied boundary segment coordinate strings.

(8) Facilities to aggregate geographic individuals and associated attributes into larger units.

(9) Facilities to disaggregate similarly on the basis of polygon overlay techniques and on the basis of specified rules for allocation of attributes.

(10) An ability to map selected attributes in various ways at different scales and with a choice of symbolism.

(11) Facilities for modifying coordinate data from one coordinate system to another given control points, for example digitising coordinates into National Grid coordinates, and for transforming coordinate data on the basis of mathematical expressions, for example from digitising coordinates into a specified map projection.

(12) Internal checking of data for consistancy, for example checking boundary segment end points to ensure polygon closure, and internal provision for checking if requested operations are feasible given the facilities available.

And in general:

(13) The system should be reliable, easy to maintain, up-date and extend.

(14) It should have internal facilities for ensuring data security and confidentiality where necessary.

(15) It should be easy to use by people having no particular computing knowledge, preferably being implemented under a terminal-based interactive computer system.

Additional to the above points if the system is to provide for many different types of spatial data, then it is highly likely that it will need to be capable of accepting data with different types of spatial reference, for example data pertaining to points, lines and areas. The system should be capable of manipulating, analysing and mapping these, and also converting easily between each, a requirement which may entail the making of assumptions in the use of techniques such as the creation of Thiesson polygons.

1.2.2 The User Interface

"Ease of use" is one of the more important issues in the list given in section 1.2.1. The term is used here to mean simplicity, in that the user need not become familiar with esoteric jargon, or require detailed knowledge of internal functioning, and also to imply that procedures should be non-tedious. Interactive statistical packages, such as MIDAS, allow the user to perform otherwise time-consuming multi-variate analyses simply by typing a keyword, such as "CLUSTER", along with selected variable labels. The very basic concept that the computer is a tool for making otherwise tedious manual tasks much easier, or even permitting tasks to be carried out which otherwise could not be contemplated, is often forgotten. An extreme example of the latter is the "travelling salesman" problem which requires the exploring of thousands of combinations of alternative routes to find the optimum. It may be argued, for example, that all users should write their own programs rather than use ready-written packages for the reason that packages and the algorithms they employ are of unknown reliability. However, this same arguement applies at the lowest levels where compilers or assemblers may contain serious bugs or even the machine logic may be in error! If this arguement were to be followed up, an inordinate amount of time would undoubtedly be spent on setting up the appropriate programs for the task before using them to produce the desired output, with the possible result of reducing the user´s work to relative insignificance. There is no suggestion here that data should be fed into one end of a statistical package or an information system and results received from the other without the user understanding the theory of the analysis being performed, simply that it may not be necessary for the user to understand fully the programming procedures going on in-between, provided that he has knowledge of the algorithms employed. Users have to take the efficiency and reliability of the program on trust to a certain extent, just as one must often take the reliability of any piece of manufactured equipment on trust. However, manufacturers should ideally test the reliability of their products, software or hardware, themselves and publish reliability estimates in their user manuals.

Assuming a terminal-based system, one common method of providing a simple-to-use interface between information system and user seems to be a query/reply procedure. An interfacing program is written which interrogates the user via a hierarchical system of questions, each supplying a range of possible replies. For example, a geographical information system might begin with an introductory section explaining the data and facilities that are available and how to respond to interrogation. Preliminary questions may be concerned with discovering with which section of the available data the user is interested, both in terms of variable names and type of geographic individuals. The user is presented with a list of labelled options, and replies by typing in the appropriate label or labels, ideally in free format using a seperator such as a blank. Selection of options will result in the calling up of data retrieval routines, and routines which link to statistical packages or to graphics packages, according to the user´s specification. This query process continues until the precise user requirements are obtained, causing the information system to initiate production of these, preferably simultaneously or otherwise after the interrogation has ceased. In a sophisticated system, the user should be able to intervene if the process goes wrong, and either go back to previous stages or abort the job altogether. "Help" and "news" facilities may be introduced, as well as some method of obtaining user feedback so that complaints and recommendations may be left for those responsible for the up-keep of the system.

From a user´s point of view it would, of course, be more desirable to be able to type in a series of statements defining exact reqirements rather than answer a lengthy set of questions with options, but this would involve a lot more work for the systems programmer in order to ensure all possible forms of requests were understood. However, recent research into "natural language" data base management systems (for example, Wallace, 1981 and W.Pitkin, 1981) has proved relatively successful. These enable the user to phrase a question in the way in which he would naturally ask it. The system parses the sentence and identifies the words by reference to a directory. Problems which arise are mainly from coping with ambiguous or nonsense questions.

In some ways, the interrogation procedure has its advantages for the naive user in that it may help to structure his or her requirements in a logical manner. However, the practised user may find repeated journeys through the interrogation very tiresome and would benefit from an alternative interface to which he could enter commands and queries phrased using vocabulary familiar to the system and which would enable direct access to any of the systems capabilities without having to pass through intermediate steps of the interrogation procedure. Interrogation is clearly not an efficient method to use when operating in batch mode; an alternative command mode is necessary.

1.2.3 Implementation And Software Support

Geographical information systems may range from those which have been written from basics for a specific purpose, including graphical and statistical facilities - often with very specific hardware devices in mind - to those which depend strongly on an already operating general-purpose main-frame computer which provides graphical and statistical subroutine libraries and packages. The advantages of creating the latter kind of system are clear since much of the "lower level" programming has already been done. The remainder of the work consists mainly of linking those relevant libraries and packages together, providing a user interface program which calls up the appropriate linking routines as the user selects options, and writing additional data manipulation programs which the computer system does not already provide. Most mainframe computers provide graphics software in the form of subroutine libraries such as GINO and GHOST which include routines for tasks such as drawing and scaling axes, drawing various projections of surfaces and contouring. Graphics software for choropleth, proportional symbol and similar thematic mapping generally comes in package form due to the complexity of coping with zone boundaries, zone shading and the drawing of proportional symbols. However, the Geographic Algorithms Group (GAG) have recently developed a library of over 160 subroutines, able to carry out a variety of data validation, manipulation and mapping tasks (Elsey,Glynn and Milne, 1981). GIMMS ("Geographic Information Mapping and Manipulation System" - discussed in section 3.3) is an example of a mapping and data management package which requires no programming by the user, but simply responds to a series of commands in a similar way to a statistical package like SPSS. Mention should be made of SYMAP which is a relatively long-established mapping package which converts vector encoded data input into raster format suitable for producing Thiesson polygons, choropleth and contour maps on a line printer (Schmitt and Zafft, 1975).

Just as it is preferable to be able to edit and manipulate data values interactively, it is also preferable to deal with graphical representations of coordinate data interactively. Interactive graphics is a fast developing aspect of computing and there is much recent literature on the subject (Newman and Sproull, 1980).

Mapping packages, such as GIMMS, and more recent versions of MIDAS provide a facility for interactive graphics where the user is able to edit coordinate data not just by looking at strings of coordinate figures, but by viewing and interacting with the image described by those coordinates using a device such as a Tektronix screen. With the use of cursor cross-hairs or a light pen, the user may annotate a map, remove and replace sections of the map, enlarge or reduce the map as required. The final version may be saved and eventually sent to a hard copy device for plotting. Many of these libraries and packages are designed to be device-independant. This is a desirable quality for an information system as a whole, unless it is particularly special-purpose, since it is likely that different sites may want to implement it on their own particular hardware.

The software facilities not generally already available on mainframe computers tend to be those dealing specifically with the management of the spatial aspect of the data. These include facilities for information retrieval on the basis of spatial delimiters, as expressed under the desiderata in section 1.2.1. For example, point-in-polygons routines which enable point data to be retrieved on the basis of presence within a given area defined by a coordinate string, or polygon overlay methods for reallocation of attribute data to a different set of geographic areas. Recent recognition of the need for portable spatial data management routines like these has encouraged various groups to supply them either as a library of subroutines (GAG) or included in a mapping package (GIMMS). These routines are probably as important for those concerned with spatial data as conventional statistical routines, if not more so, as the brief discussion of exploratory methods attempted to show in section 1.2.1.

1.2.4 Geographical Data Organisation And Encoding Techniques

In order to handle the various data manipulation procedures necessary for a geographical information system, an efficient method of structuring data within files is essential. There are a number of commonly used methods by which data base management systems structure data. INGRES is a data base management system which runs on the UNIX operating system and which utilises the relational model, structuring data as a series of case-by-variable matrices or "relations" which cross-reference one-another. A geographic retrieval system, GEO-QUEL, has been written to link with INGRES and structures coordinate data in relational form (Go, Stonebraker and Williams, 1975). Much literature has been written on the subject of data base organisation and it is beyond the scope of this paper to discuss it in detail. However, some important points with regard to the management of spatial data in particular will be highlighted.

Rhind(1976) emphasises the need for topological, geographic and attribute directories if all retrieval operations for different types of spatial data are to be adequately handled on existing multi-purpose non-intelligent hardware. He explains the advantages in storing sophisticated descriptors along with the data so that the system may recognise the nature of the data it is dealing with. In this way the system is able to find any available data of a specified type for a given task, to convert from one type of data representation to another, to report back to the user if the requested task is not valid for the specified data, and to carry out efficient error reporting. A good classification scheme must be adopted in order to define such data descriptors and a list of suitable criteria on which to base such a classification is given by the author. He gives an example of a possible data structure, showing how relationships between the elements are explicitly expressed.

Where the dimensions and extent of the geographic individuals on which certain attributes are based are explicitly expressed within the data, the structure is known as "vector" encoding. Some relationships may be implicit within the organisation of spatial data, for example values for a single variable may be held as an n by m matrix within a file, the columns (m) and the rows (n) themselves representing the plane onto which the values are mapped. Thus the contiguities are built in to the structure itself. This type of data storage is known as "raster" encoding. Peuquet (1979) discusses the advantages of raster storage in some detail. Some of the more important points are outlined here.

Data stored in raster form have advantages in implicit contiguity, the fact that two-dimensional arrays are easy to deal with for most programming languages and now may be even more easily manipulated by parallel processors such as the DAP (Digital Array Processor). However, disadvantages are that the data are of a fixed, pre-defined resolution, the resolution units are purely abstract in that they bear little resemblance to true patterns in the data, and there may be a lot of data redundancy where map pixels hold no information. Vector encoding, however, means little data redundancy and the spatial resolution is self-adjusting.

The two methods of encoding are not necessarily directly competitive. One method may be more suitable in some situations than the other, and vice versa. If the mapping is to be of the dasymmetric type where the data themselves define the area boundaries, raster encoding may be more suitable. For example, if a regular grid of pits had been dug in an area to determine variation in soil quality, it would be a simple matter to arrange the attributes in files where each postion in the n by m matrix represented a grid square. A package such as SYMAP could easily produce a line-printer map for each soil attribute using a specified gray scale, (See Webster, 1977). However, where region boundaries are previously known and need to be stored for choropleth mapping of a variety of attributes, vector encoding is likely to be more suitable. In some cases it may be necessary to have the facility to convert data stored in raster form to vector or vice-versa. For example, a choropleth map may be stored as a set of coordinates representing region boundaries, followed by a set of shading values, each associated with a region. If the available graphics terminal for displaying the map is a raster device, it will be necessary to convert the map representation from vector to raster form. The following paragraphs are concerned with methods of vector encoding, relevant to the subject matter of this study.

Where the geographic individual to which the attribute values refer is a point in a plane, the information required to define this is simply an x and y coordinate which may appear in a file record along with an index value such as a name, linking it to its associated attribute file record, and a value explicitly stating that it is a point. Where the geographic individual is a line then this is usually defined by a coordinate string of at least two coordinate pairs. Preferably, as Rhind (1976) shows, stored in association should be the minimum and maximum x and y coordinates, the number of points defining the line and a descriptor stating that it is a line segment. As much helpful additional information defining the extent and properties of a geographic individual should be stored as is necessary in relation to predicted uses.

To define the extent of a geographic individual which is an area requires storing a series of coordinates which represent its boundary outline. Past methods of doing this have introduced a large amount of duplication, for example SYMAP requires each polygon to be named and its entire boundary to be defined by a coodinate string which begins and ends with an identical coordinate pair. Where areas are adjacent to one another and share common boundaries, the extent of the duplication is considerable. This duplication of boundary description may cause the appearance of "sliver lines" in plotted maps where two coordinate strings defining the same boundary are not identical in position. Attempts to reduce the degree of redundancy in the data have been made by approaching the problem, not as a set of contiguous areas, but as a boundary network. The DIME (Dual Independant Multiple Encoding) project method consists of taking each straight line section of the network in turn and recording its start and finish node numbers, the name of the polygon to the left of the section and the name of the polygon to the right. Clearly, this pre-supposes a rule for direction of coordinate recording. Node coordinates are stored in a separate file. (See Corbett and Farnsworth, 1972).

Following the ideas initiated by the DIME project, a program called POLYVRT was developed by Peuker and Chrisman (1975). They saw little need for taking each straight line section in turn since it makes generalisation or increasing the resolution needlessly difficult. The POLYVRT method takes a section of the network boundary between two polygon vertices or network junction points as being a logical topographic unit, named a "chain". Each chain has a code number, each polygon a name and each vertex a node number. The coordinate pairs defining the chain, the start and finish node numbers, the polygon names to right and left of the chain and the chain code are all stored in a series of related files. GIMMS uses the same encoding method as POLYVRT, but calls the chains "segments". The major difference between the two systems is that GIMMS requires only the left and right polygon names for each segment followed by its associated coordinate string. The package then creates the polygon itself by comparing names and joining segments on a proximity of end-points criterion. In this way, storage space is reduced but time and effort for creating the polygons is considerably more than for POLYVRT which only needs to use "look-up" procedures. However, once GIMMS has formed the polygons from the segments it accepts as input, a "polygon" file is written which can be stored for future GIMMS map runs so that, providing no changes are to be made to the segments, the polygon-creating routines need only be used once. GIMMS will also cope with polygon aggregation hierarchies, where the polygons of each aggregation level have a code as well as their individual labels, (see section 3.3).

The situation becomes slightly more complex where there are several overlapping boundary networks to be considered. If each network were to be seperately encoded, the only connection between them expressed within the data would be by laborious searching through the coordinate values, making the task of determining the spatial relationships between the networks a very tedious one. To include each network within one encoding system, the Australian South Coast Project (B.Cook, 1979) defined a "least common geographic unit" (LGCU) which is a polygon uncut by further partitioning, formed by the overlapping networks. A portion of the boundary of this LGCU between a vertex or junction point is regarded as a chain so that the boundary between two polygons for a given network system is made up of "chain groups". The main drawback of this encoding method is that the original polygons, that is the polygons for each of the overlapping networks are not preserved within the stored data. An even more complex situation arises where the geographic individuals concerned are volumes, such as may be faced when dealing with geological formations, and the common boundaries between two units are not lines but planes. Whatever the dimensionality of the geographic individuals, it is essential to express within the data the spatial associations between them, since it is precisely these which are the concern of most investigations. The encoding technique employed to do this should be efficient in terms of reducing data redundancy, while permitting easy extraction and manipulation of spatial units. The GIMMS and POLYVRT solution seems to be the most efficient vector encoding method for two-dimensional geographic individuals. I have no knowledge of any practical attempt to define an encoding technique for three-dimensional individuals.

References

COOK, B. (1979) "Land Use on the South Coast of New South Wales" Chapter 5.

CORBETT, J. and FARNSWORTH, G. (1979) "Theoretical Basis of DIME in U.S. Bureau of the Census". In U.S. Department of Commerce Conference Proceedings. International DIME Colloquium, Washington D.C..

COX N. (1978) "Exploratory Data Analysis for Geographers". Journal of Geography in Higher Education. Volume 2(2) p51-54.

ELSEY, T. GLYNN, S. and MILNE W. (1980) "An Algorithms Library for Spatial Data" IUCC Bulletin Volume 2(3).

GO, A. , STONEBRAKER, M. and WILLIAMS, C. (1975) "An Approach to Implementing a Geo-data System". Presented at the 16th IFIP/IAG Data Base Workshop.

HARRISON J.C. (1979) "The LAMIS System". Chartered Land Surveyor/Mineral Surveyor. Volume 1(2)

NEWMAN, W.M. and SPROULL, R.F. (1980) "Principles of Interactive Computer Graphics" 2nd edition, McGraw-Hill.

PEUKER, T.K. and CHRISMAN, N. (1975) "Cartographic Data Structures". American Cartographer. Volume 2, p55-69.

PEUQUET, D. (1979) "Raster Processing: An Alternative Approach to Implementing a Geo-data System". American Cartographer. Volume 6(2), p129-139.

PITKIN, W. (1981) "Panda - A Data Base Query Language for Ordinary People" Paper presented at the IUCC Conference.

RHIND, D.W. (1976) "Towards Universal Intelligent and Usable Automated Cartographic Systems". ITC Journal. Number 5, p515-545.

SCHMITT, A.H. and ZAFFT, W.A. (1975) "Programs of the Harvard University Laboratory for Computer Graphics and Spatial Analysis". In "Display and Analysis of Spatial Data" by Davis, J.C. and McCollagh, M.J.

(eds.), Wiley.

THOMPSON, C.N. (1979) "The Ordnance Survey Topographic Data Base: Concepts for the 1980´s". Paper presented at the _2nd United Nations Regional Conference for the Americas_, Mexico City.

TUKEY, W. (1977) "Exploratory Data Analysis". Addison-Wesley.

WALLACE, M. (1981) "Natural Language Access to a Relational Data Base". Paper presented at the _IUCC Conference_.

WEBSTER, R. (1977) "Quantitative and Numerical Methods in Soil Classification and Survey". Oxford: Clarendon Press.

CHAPTER 2

EEC DATA AND USER REQUIREMENTS

2.1 USER REQUIREMENTS

The unification of geographical areas under a common decision-making body, for whatever purpose, generally necessitates the compilation of information relevant to the area as a whole, and the European Economic Commission is no exception. The availability of Europe-wide social and economic statistics is integral to the development of the Community, allowing the monitering of its progress towards planning for its future. However, it is not sufficient simply to compile whatever data becomes available since the result will almost certainly be a collection of isolated sets of unrelated statistics, particularly where different countries under different administrative regimes are concerned. There are certain properties of comparability and consistancy which the data must show, over the entire EEC, to permit fruitful research and effective planning to be accomplished.

The process of regional planning automatically involves drawing comparisons, both on a spatial and temporal basis. In order to delineate areas of high unemployment, poor housing, or ageing population, for example, the researcher must be provided with data which is associated with <u>comparable</u> geographic units (or geographic "individuals"). The term "comparable" must, however, be qualified, and for socio-economic spatial data, which will be most relevant to EEC regional planning, it is usually sufficient for regions to be similar in terms of areal extent or population figures. Geographers have been aware for some time of the often misleading effects produced by the mapping of variables for different sizes of geographic individual, particularly where ratio data are concerned (Visvalingham (1978), Robinson, Sale and Morrison (1979)).

The problems inherent in the analysis of incomplete data sets are well-known to social scientists, and those concerned with using EEC regional data would be extremely fortunate if no such difficulties existed. Statistical routines in most well-known computer packages, such as SPSS, are able to cope with missing data to a certain extent, but problems can still be great when using more sophisticated statistical techniques. There must, therefore, be some pressure to compile complete data sets where possible. Features which may affect data availability include psychological reasons, for example, sensitive ethnic questions or income questions, or administrative reasons such as changes in boundary structure of the geographic individuals to which the variables relate.

It is often necessary to compare data pertaining to different networks of areas. This can arise where the data has been compiled by seperate authorities for different geographic individuals, or where boundaries of geographic individuals have changed over time. Boundary changes are of particular relevance to the U.K. which had a major re-shuffle of local government areas in 1974 and 1975, making provision of time-comparable data a difficult task, (Tanenbaum and Taylor, 1980). Some attempts have been made to compile data sets for pre- and post- 1974 boundaries, notably the efforts put into re-structuring the 1971 Census data sets so as to render them compatible with 1981 data. This "polygon overlay" problem is simpler when dealing with sets of areas which are regularly shaped and nest neatly, one into another, but is substantially harder if the polygons are irregular with overlapping boundaries and do not aggregate easily. Rhind (1980) explains how he has avoided overlay problems where possible by aggregating the more detailed data to fit within the coarser networks. Attempts were made to allocate enumeration districts (EDs) into Employment Office Areas for the U.K. Northern region, on the basis of ED centroids. This method can result in the mis-allocation of much information by failing to take account of the shape of the ED. If extra information is available, for example about the distribution of population within each ED, it may be possible, for example where a smaller area straddles the boundary of a larger one, to allocate information more accurately.

One-to-one matching of geographic individuals where the attributes reference supposedly the same point or area on the ground, also has its problems. Two sets of variables may reference a point in a different way, for example, one by grid reference, one by postal address. Several good address-matching programs now exist to link up such data sets (S.Openshaw, pers.comm.). Additional problems arise where the characteristics of the geographic individual may change but its reference name in the data set is the same. Data may apparently be available for a particular named geographic individual over a long time span, but the fact that the name remains consistant throughout is no guarentee that its boundaries have not undergone substantial re-positioning.

The type of geographic individual for which data is available is also important to the user. For comparability, there are obvious reasons for using geographic individuals of the same type. It is also important to ensure that those individuals chosen as a basis for analysis of certain variables really do reflect variations in those variables. For example, if the purpose of a study was to discover regional differences in employment figures, it would hardly be illuminating to use geographic areas based upon, say, height above sea level. The boundaries of one network might fall neatly across the regions of the other, concealing the true disparities which may exist in employment levels across the country. This may be thought of in terms of the delimiting boundary network rather than the geographic areas themselves, where the placement of the network for analysis must be such that it coincides with the position of the real network as defined by the variable itself. In practice, some compromise often has to be made between the spatial patterns given by the different variables and also the geographic individuals for which data is already available.

Resolution is important in this context also. If the resolution of the geographic individuals chosen for analysis is not close to that displayed by the variable in question, misleading results could well follow. A strong correlation between two variables at one level of aggregation may be weak and insignificant at another. Yule and Kendall (1950) and Openshaw (1977) have shown how it is possible to find a whole range of correlations between two variables simply by altering the geographical resolution at which they are studied. Such ecological correlations may be accounted for in analysis by using analysis of variance for chosing the most appropriate scale, in terms of the true variations in the data, or testing for spatial autocorrelation (Cliff and Ord, 1975). However, ideally data should be available for as many degrees of aggregation as possible so that there is no temptation to generalise findings at one level for all levels.

The data themselves must be consistant in the <u>definitions</u> of their variables from region to region. Norris, Rhind and Hudson (1980) cite several examples of <u>within</u> country problems along these lines, where figures for supposedly the same variable differ drastically. If there is a problem within countries, it will be even more emphasised across national and cultural boundaries. Even the most apparently innocuous variable, for example, "total population", may be defined inconsistantly.

There have been some attempts to encourage data collection to be undertaken on the basis of some small, regular geographic unit such as the 1 kilometre grid square, which has little association with the pattern of the variables being studied. This may sound contradictory to previous discussion concerning coincidence of boundary networks with true variable patterns, but this is not so, since 1 kilometre grid squares may readily be aggregated up in such a way as to fit the requirements of the researcher and his data. Each grid square unit is exactly the same size and shape and is therefore directly comparable with all others, such a grid is unlikely to be subject to boundary change pressures over time because of its irrelevance to administrative regions, and there is a ready made continuous hierarchy of aggregation levels. Movement between aggregation levels should cause little difficulty and therefore encourages the user to study data at different resolutions to account for ecological correlations.

In order to preserve the utility of historic and current data sets, the EEC data collection authorities have to rely to a large extent on existing geographic individuals as defined by each individual country for its own purposes. Despite these constraints, much effort should be made to encourage comparability in geographic individuals and consistancy in variable definition by providing standards towards which each country may work for data provision. Clearly, it is unrealistic to expect each country to change its entire data collection basis to comply with these standards, but it may be persuaded, in its own interest as part of the EEC, to collect data in such a way as to allow comparison with its fellow members and also to attempt to re-structure past data as far as possible.

2.2 EEC DATA PROVISION - CURRENT AVAILABILITY AND RECENT DEVELOPMENTS

The Statistical Office of the European Communities, based in Luxembourg, is set up to gather and publish EEC data in a form suitable for distribution to all potential users, but particularly to the Community´s policy-makers, providing them with essential information on which to base planning decisions and to allocate EEC funds. The result of these efforts is an extremely large and growing statistical data base which is made generally available via EUROSTAT publications.

The task alloted to EUROSTAT is far from straightforward. In most cases, the organisation is dependent upon the work of national statistical bodies for data provision, whose methods of collection and relative stages of development may widely differ. EUROSTAT therefore aims also to encourage the coordination of these national bodies, with the hope of acheiving some standardisation in technique between countries. EUROSTAT has managed to establish fairly strong links with the various statistical institutes of the nine member states; some seventy different statistical groups now meet in Luxembourg. The institutes themselves have in fact needed little encouragement from EUROSTAT, many showing great enthusiasm for improving communications between one another. This has led to the organisation of a number of international meetings which in turn have initiated the formation of various working parties whose general aim is one of facilitating information interchange and promoting information consolidation. (See La Cour, 1980 and Mesnage, 1980).

In attempts to lay-down some standardisation of geographic individuals over the entire EEC, a number of proposed regionalisations have been put forward, notably that of Lorentzen and Rokkan, (1976). EUROSTAT finally developed a four-tier hierarchy of regionalisation levels, ranging from boundaries of member states, through "Level I" and "Level II" down to the "Level III" (French department or U.K. county boundaries). Socio-economic data are currently available from EUROSTAT in book form for Levels I and II over all nine member states.

At Bergen in 1977, the Norwegian Social Sciences Data Services (NSD) arranged a conference for European research workers and representatives of data services and centres of cartography to discuss the establishment of a joint data bank for thematic mapping of regional variations in Europe. By this time, the NSD, a data archive organisation, had already developed facilities for computer mapping of variations across the Level I regions of Europe (Lorentzen and Rokkan, 1976) using the programs POLYVRT and CALFORM, both developed by the Laboratory for Computer Graphics and Spatial Analysis at Harvard University. Papers presented at the conference by Lorentzen and Rokkan put forward ambitious ideas for the development of a multi-level data bank for computer mapping of regional data for Europe. Their scheme was to be based on the EEC´s regionalisation for six member states, but extended to the now nine member states and also to other countries in Western Europe with the possibility of using a fifth level. Their plans were to work in fairly close collaboration with the EEC to avoid any conflict or duplication of efforts. The primary sources of the data were to be recent EEC statistics as compiled by EUROSTAT and the Handley file. The latter is a combination of regional EEC socio-economic data and EEC survey data which Handley used to identify problem regions in six European member states in 1975. (See Handley, 1976)

At the conclusion of the Bergen meeting, it was agreed to establish a European Working Group on Databases for Regional Analysis, the priorities of which were outlined as follows:

(a) To promote efforts to build up data base systems in each country at the lowest possible levels of aggregation and to link these with facilities for automated cartography.

(b) To prepare inventories of available information to be circulated throughout the network of statistical services in Europe.

(c) To extract from the available data files a number of cross-nationally compatible packages of data for comparative research.

It was agreed that, during 1978 and 1979, the working group was to produce a set of computer maps of various socio-economic data for Levels II and III to parallel the efforts of the Council of Europe in promoting maps of levels of urbanisation across Europe. The objective was to demonstrate the superiority of computer mapping as a tool for regional analysis so that maps of multi-variate groupings as well as individual variables were regarded as important.

In practice, this project took a longer time to get underway than first anticipated and 1978 and 1979 were taken up with discovering where and in what form the sources of data were, and their relative compatibilities. Sande and Rokkan (1980) drew up a table showing the properties of some of the national data sets available in the countries of Western Europe.

The International Federation of Data Organisations (IFDO) arranged a conference to be held in Turin in the spring of 1980 - a Symposium on the Development of Joint Data Bases for Regional Analysis and Computer Cartography. Participation was restricted to those researchers and administrators willing to prepare reports on the development and utilisation of computerised data bases at local/regional levels for their own countries or across several countries. Various international bodies were also represented, including EUROSTAT. The purpose was to review each country´s progress in the field of data base development and to give greater impetus to the coordination of efforts for the establishment of a joint data base of regional statictics for Europe. The Handley file, by then extended to the nine member states of the EEC, was seen to be a major source of data for regional statistics at the Level II degree of aggregation. IFDO hoped to establish a European regional data base for Level III geographical units also, the lowest level of aggregation possible being the most desirable. The time span for both levels was initially to be from 1945/50 to the present. The word "European" in this context was not seen to be limited to the EEC member states but to extend to as many European nations as possible. However, the definition of a fixed area within Europe was thought necessary, at least initially, for various technical reasons and so as to concentrate effort. Expansion could be made at a later stage. IFDO saw the overall goal to be "a jointly maintained infrastructure service for comparitive European research".

Nineteen country reports and thirteen other papers were presented at the symposium. All these were published on a two-volume set by CSI. IFDO clarified its position as a coordinator and reference body for contacts, emphasising that it did not intend to be the only agency for carrying out the specified aims.

Most papers given were reports on specific projects, their problems and utilities, by various representatives from different countries. A few of these projects involved cross-national regional comparisons, for example, Gorzelak´s study of regional differences in living standards between Czechoslovakia, France, Japan, Poland, Spain and GDR. He found it necessary to narrow his number of variables for analysis down to fifteen; constraints are always provided by that geographic unit for which the least amount of data are available. He also found that several assumptions had to be accepted which could not be confirmed by reality, thus introducing a certain amount of likely error into his analysis.

Norris, Rhind and Hudson (1980) delivered their paper "Alchemy in Action? Regional Data Bases, Statistical Analysis and Cartography in a Computer Environment" in which they expressed great concern over the apparent naivity in approach of many people involved in setting up the European data base. Reference is made in particular to the Lorentzen and Rokkan regionalisation paper (1976) and its failure to point out the traps which await the unwary researcher. Lorentzen and Rokkans´ major criteria in the selection of appropriate geographic indiviuals appeared to be that aggregation between levels should be easy and that the area sizes should be convenient for mapping. Norris et al make it clear that these are by no means the only issues, and outline the other relevant considerations discussed in section 2.1. Specific to the EEC, these were defined as follows:

(a) Data must be available for every geographic individual from 1945 to the present and variables must be defined consistantly between countries.

(b) The within-country variation of the area or population of these geographic individuals should be no more than the between-country variation.

(c) These geographic individuals should be based upon "real" variations in the variables.

(d) Analysis of the data must be possible at different levels of spatial aggregation.

The authors show how, apart from Belgium, the difference between intra country variation and inter country variation in the Lorentzen and Rokkan regionalistion is at a relatively acceptable level, and they agree that it would be hard to find an improvement on the system suggested.

Tanenbaum and Taylor (1980) at the Turin conference described the problems of a similar nature encountered when dealing with U.K. data. The situation as it stands for the U.K. is that data for the Lorentzen and Rokkan Level III prior to 1974 or for a finer spatial resolution for any time period are not readily available and may indeed prove difficult, or at least time-consuming, to bring together.

The outcome of the IFDO meeting in Turin was to specify more closely those ideas which had been discussed in Bergen, to draw up a more detailed plan for future development, and to delegate responsibility for seeing these plans through. The concept of a joint data base for European regional analysis emerged as a practicable objective from the information given in the country reports. NSD in Bergen was given the task of coordinating the agglomeration of data files for Level III units. Much information was seen to exist at Level IV also, but it was clear that steps towards consolidating information for the joint data base were much more complex than for higher aggregation levels, so IFDO decided not to set up any well-defined plans but simply to nominate Michael Aitkin and Guido Martinotti (president of IFDO) as preliminary coordinators until more firmly established contacts between nations had been made. Roy Drewitt at the London School of Economics and Guido Martinotti were asked to recommend methods of discovering priorities for areas of comparitive research. David Rhind at Durham University was asked to liaise with the IFDO Working Group on Computer Cartography towards the production of computer drawn maps of Europe for the desired levels of aggregation. IFDO suggested that drawing up a list of available coordinate data bases and a list of those centres able to provide facilities for automated cartography would be fruitful.

2.3 PROPOSAL FOR FUTURE DEVELOPMENTS, AND AIMS OF THIS STUDY

The present status-quo as far as the joint data base is concerned is the result of decisions taken at the Turin conference, leaving various academics and members of statistical institutes with responsibilities for different facets of the project. Despite these encouraging developments and the general enthusiasm expressed by many European countries, there seems as yet no concrete structure on which to build and no well-defined final objective. IFDO looks towards a "jointly maintained infrastructure", a term which could cover a whole range of possible outcomes. Although there is still a lot of work to be carried out by each country in compiling data for the specified geographic units, it is still necessary at this early stage to clarify the picture for the later stages. A well-known pitfall for researchers is the discovery that their data has been collected in a way far from ideal for the kind of analysis required, or that their data are not in a form suitable for management and analysis by computer. Hopefully the NSD will remain aware of the problems which exist, some of which were outlined by papers delivered at the Turin conference, when coordinating the formation of a data base for Level III statistics.

In an attempt to look ahead, therefore, I shall try to present a fairly clear picture of what I regard should be the objective in mind of those invoIved in providing and/or using regional statistics for Europe. "Europe" in this case means the European Economic Commission to which I have decided to confine my work. The reasons are partly technical, namely the availability of data from EUROSTAT and of pre-defined geographic units for comparison between regions, so that I can adequately demonstrate my ideas in the following chapter. However, the main reason is because the EEC is a political community in its own right, which has its own parliament and so can plan as a whole for its future. I therefore envisage the development of a joint data base as being useful not only for researchers as defined by IFDO, but also for the planners and administrators of the EEC.

At present, EUROSTAT provides information in the form of a printed publication. Clearly, for any user wishing to carry out some detailed analysis, this is likely to be inadequate, since he will almost certainly want to make use of computing facilites. For this reason and for reasons of storage and access efficiency, EUROSTAT intends to compile all available regional data in computer-compatible form and to supply it on magnetic tape. However, as section 1.2.2 of this study attempts to demonstrate, potential users of the regional data base will not necessarily know about computing in general, and in theory should not need to in any depth since it could represent considerable digression from the true nature of their work. The computer should ideally act purely as a tool for these users, for example in enabling efficient information access for administrators or in aiding regional strategy evaluation for planners. In short, the requirement is for an information system able to cope with management of the EEC regional data files, able to provide data analysis and mapping facilities, able to provide formatted output, and able to present a "friendly" interface to users who should not require knowledge about the internal structure of the system.

The time and work required for the writing of such an information system for the regional data of the European Economic Community should not be underestimated, even if the design were to include existing operating system software. Because of time constraints, it was considered impracticable to embark upon such a task for the purposes of this dissertation. The aim of this study is therfore to <u>test the feasibility</u> of creating an EEC information system given the difficulties and constraints which might be involved, and given the current availability of suitable software. It aims also to demonstrate the usefulness of such a system for selected users of the EEC regional data.

References

CLIFF, A.D. and ORD, J.K. (1975) "Model Building and the Analysis of Spatial Patterns in Human Geography". Journal of the Royal Statistical Society, series B, volume 37, p297-347.

HANDLEY, (1976) "A File on the Regions of 5 EEC Countries". European Political Data, number 18.

LA COUR, A. (1980) "Policy and Recent Developments in EUROSTAT - Methods of Disseminating Statistics". European Political Data, number 35, p7-16.

LORENTZEN, J. and ROKKAN, S. (1976) "A Multi-level Data Base for Computer Mapping of Regional Variations in Western Europe". European Political Data, number 19.

MESNAGE, M. (1980) "Dissemination of Statistics using Computer Facilities in the EUROSTAT". European Political Data, number 35.

NORRIS, P, RHIND, D.W. and HUDSON, R. (1980) "Alchemy in Action? Regional Data Bases, Statistical Analysis and Cartography in a Computer Environment". Paper delivered at the IFDO Conference, Turin, March.

OPENSHAW, S. (1977) "A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning and Spatial Modelling". Trans: Institute of British Geographers. New series, p459-472.

ROBINSON, A., SALE, R.D., and MORRISON, J.L. (1979) "Elements of Cartography", Wiley, 4th edition.

SANDE, T. and ROKKAN, S. (1980) European Political Data, number 34.

TANENBAUM, E. and TAYLOR, M. (1980) Paper delivered at the IFDO Conference, Turin, March.

VISVALINGHAM, M. (1978) "The Signed Chi-squared Measure for Mapping" Cartographic Journal. Volume 15(2), p93-98.

VISVALINGHAM, M. and DEWDNEY, J.C. (1978) "The effects of the size of

Areal Units on Ratio and Chi-square Mapping". Census Research Unit Working Paper, number 10, University of Durham.

YULE, G.U. and KENDALL, M.G. (1950) "An Introduction to the Theory of Statistics". Charles Griffin.

CHAPTER 3

THE APPLICATION OF SELECTED SOFTWARE TO EEC REGIONAL DATA

3.1 INTRODUCTION

This chapter describes the practical work which forms the core of this study. The intention is to demonstrate how a workable information system, suitable for users of European regional data, could be formed from the basis of a number of existing packages and operating system utilities. EEC Level II boundaries were digitised, and selected 1977 socio-economic data transferred onto magnetic disk in order to provide some concrete examples of the potential of such a system. Particular emphasis has been placed on mapping the regional data. The importance of viewing spatial data in its context has been explained in section 1.2.1; this has particular relevance for those concerned with EEC regional inequalities.

3.2 THE COMPUTER AND OPERATING SYSTEM

In practice, work had to be confined to making use of software implemented on the IBM 370/168 computer at Newcastle University. This machine has direct lines to Durham University and to Newcastle Polytechnic and is known as NUMAC (Northern Universities Multiple Access Computer). The operating system implemented at NUMAC is MTS (Michigan Terminal System), software written by the Computing Centre staff at the University of Michigan and the University of British Colombia, who detected a demand for a powerful, user-friendly, interactive operating system for an IBM machine, in the mid 1960´s.

MTS is designed primarily for interactive use, but it also makes batch computing an easy task. Commands are familiar English language statements with a simple syntax, and tend to consist of a single command word with both mandatory and optional parameters. Error correction procedures are sophisticated in that they are easily understood, attempt to direct the user to the correct path, and are capable of coping with spelling mistakes to a limited extent. Using MTS, a user may store data files on public disc space so that they are immediatelty accessible, but so that adequate protection and confidentiality is ensured by specifying read or write access to other selected users as well as himself, if required. By default, only the user whose number the files were created under has read or write access to them, the user´s number itself being protected by a password. Disc files are set up as "line files" to facilitate information management for the user. They appear as a set of numbered lines; each line can be a maximum of 32767 characters in length. A file may be listed or submitted to programs in its entirety or partially by specifying line number ranges. Files may also be concatenated for similar operations, simply by specifying several file names interspersed with plus signs.

The disadvantages of running such a user-friendly operating system, which some sophisticated users feel outweigh the advantages, is that the machine response can become almost unacceptably slow when the system is heavily used. Structuring information in a form suitable for a user is very often inefficient for the machine in terms of time and space constraints. For example, structuring disc files as line files with variable length records means storing beginning and end of record markers, and searching for these whenever the file is manipulated in some way.

Parsing commands, error correction, accounting and file protection checks are all required to load and run large programs, and carry out numerous disc-access procedures - all of which take up much machine time. Interactive systems also tend to "clog-up" more easily than batch systems, since the former requires immediate action by the computer, but the latter can queue jobs until the machine is ready to accept them and so the work is spread out more evenly throughout the day.

In addition to the usual programming language compilers residing on the computer system, there are various subroutine libraries both for graphics and numerical analysis, and a number of graphics and statistical packages, many of which are simple to use and are well documented. NUMAC, like most other University Computer Centres, has the difficult task of trying to provide an efficient service for all types of users with widely differing requirements, from those who need powerful yet high precision arithmetic operations with fast turnaround time, to those who care more about file management and data editing capabilities. For an information system, the efficiency of the latter is, of course, extremely important.

MTS has a powerful editor whose utilities include global search and replace procedures, easy within-line changes, setting line and column ranges, and transferring or copying lines from one file to another. Editor executive procedures may also be set up which can considerably reduce the tedium of editing large data files. A screen editor has recently been implemented which, although probably most useful for text editing, can be far quicker and less prone to errors for some procedures than a context editor. Other useful utilities on the MTS system include sorting routines which can arrange records within files either numerically or alphabetically according to any specified column or range of columns.

All these system utilities, along with the mapping package GIMMS, the statistical packages MIDAS and CLUSTAN, and a number of self-written FORTRAN programs for file management and data formatting were used for this study.

3.3 GIMMS

The acronym GIMMS stands for Geographical Information Mapping and Manipulation System. The fact that it describes itself as an information system demonstrates the wide-ranging use of the term. Under the definitions given in section 1.2.1, it does not quite qualify as such, so that it seems more appropriate for this study to refer to it as a mapping package. The term "package" is generally applied to a collection of related programming routines which are accessed via a structured command language rather than via user-written programs. In this way the user need not know how to write programs but simply needs to become familiar with command sequences necessary for submitting data and receiving appropriate output.

GIMMS is primarily a thematic mapping package, and as such would form an important part of a geographical information system. It is particularly suited to generating choropleth maps, accepting coordinate data for region boundaries with associated attribute data as input, and producing, with the help of a good quality plotter, maps suitable for publication as output. GIMMS may be used both in batch or interactively.

In order to define region boundaries for choropleth mapping, coordinate data must be presented in the form of boundary segments, as explained in Chapter 1, and each must be defined by its adjacent left-hand and right-hand polygon labels followed by a string of coordinate pairs in free format. These segment specifications may be most efficiently set up in a file which can then be accessed wholesale by GIMMS. Given suitable commands, GIMMS will then structure these segments into polygons and store the latter ready for map generation. If segment end coordinates do not match, the package requires coordinate values for the polygon vertices or "nodes" as additional input and, supplied with a tolerance level, will attempt to match the segment end points with the node values to form complete polygons. Clear and helpful error messages are printed should GIMMS be unable to form an entire set of polygons, which enable the user to refer back to his or her original data and make corrections. The original segments themselves may be drawn out for error discovery.

Attribute data may also be presented as a file containing a case-by-variable matrix. Once accepted by GIMMS, individual or groups of variables can be selected for mapping, histogram production, or the creation of new variables. Choropleth map output can be rigorously defined by specifying attribute interval levels, scale, orientation and placement of the map within its frame, chosing shading symbolism from a wide range of alternatives, placement and character set for a title or other map annotations, and requesting enlarged insets of selected portions of the map. These map output specifications can be requested by the usual sequence of commands given in batch or at an ordinary VDU, but may also be submitted in a truely interactive fashion on a device such as a Tektronix. The skeleton of the map is, for example, displayed on the screen and cursor cross-hairs used to position the legend or the title, for example. This allows the user to experiment with map production until a satisfactory version is created, which can then be sent to a hard copy device.

GIMMS can also accept point and line data input for mapping. It is also able to output data and coordinates at various stages for use by other packages.

3.4 DIGITISING EEC REGIONAL BOUNDARIES

The decision to concentrate upon the 112 EEC Level II regions of the nine member states of the Community was partly determined by the availability of data for these areas, and partly by their suitability in terms of size and comparability for the purposes of this study.

The first major task to undertake - and one which occupied more time than first envisaged - was to digitise the boundaries of the Level II regions, since no such coordinate data were yet available in computer-compatible form. The digitising was based on a map produced by EUROSTAT at a scale of 1/4000000, and carried out using a large Ferranti-Cetec digitiser which was off-line and recorded information onto punch-cards. Three options for disitising method were available - point mode, stream/time mode and stream/distance mode. Point mode was chosen to keep the volume of the output within reasonable bounds and to allow for the manual selection of salient detail. The digitiser allowed a certain amount of output formatting in that each digitised coordinate recorded could be seperated from the next by a blank character or by a plus sign. Keyboard input was also accepted. The device also allowed a choice of Imperial or Metric units and a selectable level of precision down to 0.01mm (or the equivalent in Imperial Units). For this study, digitising units chosen were metric, and the level of precision was 0.1mm.

The rationale for digitising the boundaries was that adopted by POLYVRT and GIMMS, that is segment by segment. This required each polygon to be given a label and the coordinates of each polygon vertex or "node" to be recorded. Taking this approach, the output format was chosen to allow for maximum flexibility as far as was practical, so that the data could be accepted by different mapping software. It was also chosen so that coordinates and their associated labels could be recorded in a continuous fashion onto the cards, to prevent the otherwise time-wasting process of returning to the card punch after each segment was completed to start a new card. A by-product of this was that the sequence order of the cards had to be physically maintained since no card-by-card ordering was implicit in the data itself. The data items recorded for each segment were left-hand polygon label, right-hand polygon label, start node, finish node, and segment coordinate string. The digitiser was pre-programmed to separate each coordinate of a segment with a blank character. The labels, node numbers and characters used as seperators were recorded from the keyboard which was within easy reach of the digitising table. The resulting output format was as follows:

```
/left polygon label/right polygon label/start node#end node# X Y
X Y X Y X Y X Y X Y/left polygon label/right polygon label/start
node#end node# X Y X Y.......etc.
```

The total time spent digitising the boundaries amounted to 32 hours. This meant that logical break points had to be defined; these were useful from the point of view of handling the sheer quantity of cards produced and submitted to the card-reader, and from the point of view of disk file size. The obvious method was to digitise on a country-by-country basis, adopting rules for dealing with common boundaries. Since the map had to be removed from the digitising table overnight, it was also essential to ensure that the different coordinate spaces resulting from this could be easily converted one to another, or to a common coordinate system. The angle of rotation along which y values were zero was marked permanently by masking tape on the digitising table. However, it was not considered accurate enough simply to re-align the map with this line each time, so two common points on the map, subtending an angle of at least 45 degrees to the origin were digitised for each positioning of the map to enable angular corrections to be made. Before any segment digitising was carried out, the numbered nodes were digitised without moving the map position and this "node coordinate system" provided the common basis to which the other segment systems were transformed.

Prior to the digitising process, careful planning was necessary: at the time of digitising, for instance, GIMMS was unable to accept negative coordinates, so the digitiser origin was always set in the south-west corner of the map. The data recorded on punch-cards was read into disk files, country by country.

3.5 ERROR-CHECKING AND DRAWING OF REGION BOUNDARIES

A number of simple FORTRAN programs were written to carry out some initial error-checking. The accuracy of the coordinates themselves were not checked at this stage, nor were the area labels, but the correct sequence order of data items was by searching the files for the expected pattern of occurance of the seperator characters, / and #. Node numbers and area labels were checked to ensure that they were not longer than 3 and 8 characters repectively, and coordinates were counted to ensure that there was an even number for each segment. GIMMS expects coordinate data input in such a way that each segment occupies a whole number of data records, and begins with polygon labels. Each record must be no longer than 70 characters and no coordinate should be split over a record. Node numbers are not required. A copy of the raw data was therefore re-formatted using another simple character-searching program to comply with these specifications.

Initially these coordinate data were submitted to GIMMS for drawing and checking country by country. In some such cases, the boundary of the country was incomplete since common segments had been digitised and stored with the adjacent country's data. To overcome possible confusion whilst checking each individual country, straight line segments were temporarily inserted to complete the country border. To enable GIMMS to form polygons from segments, the package had to be presented with the node coordinate file from which to select "true" coordinate values of polygon vertices, and a tolerance level within which segment joining would be accepted, (see diagram overleaf). Since each country, through necessity, had been digitised in a different table coordinate space, some transformation had to be made to convert each country´s segment values into the coordinate space of the digitised nodes which was to form the common basis for the map of the entire EEC. This facility was available within the GIMMS package as a routine which required four values (two X and two Y) for two fiducial points, recorded once in each coordinate space. The segments of each country were converted into "node space" by this method, ready for polygon formation. It should be noted that this preserves the "plotting" space of the original map, that is, all maps produced without further analytical transformation merely replicate the projection of the source map.

The first country whose polygons were successfully constructed by GIMMS was France. A map of the Level II boundaries for France was produced and checked for accuracy. All other countries initially caused significant problems during the polygon construction phase, and map output from the package indicated that something had been seriously wrong during digitising. The maps for all countries except France, Eire, U.K. and Italy showed the files to contain mirror-images of the source docment in the X direction (east-west). Parts of the U.K. and Eire were grossly displaced, and the southern half of Italy along with Sardinia was folded back on itself into a mirror-image with parts of Northern Italy displaced to the east. Clearly, the coordinates for eight countries as they stood were unacceptable. The error was traced back to the digitiser itself, which had intermittently been shifting its X origin. For five of the

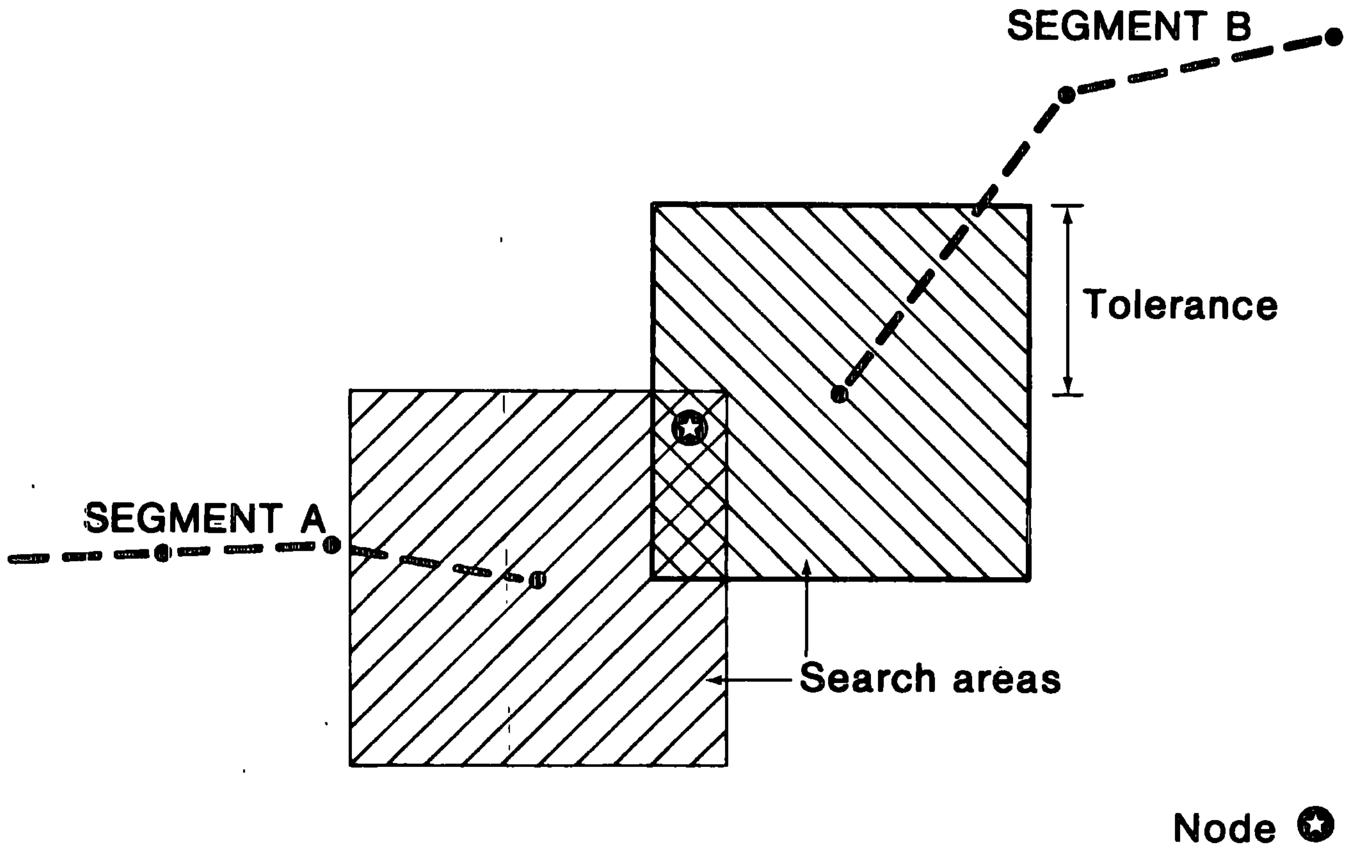


Figure 3 – 1 Diagram to demonstrate the matching of segment end-points by the mapping package, GIMMS.

countries in error, the coordinates could be corrected relatively easily, by creating and running a simple program which took negative values and added a large enough number to ensure no negative coordinates remained. The relative positions of each country with respect to one-another was unimportant at this stage, so long as the association between each country and the "node" space was maintained. The displaced segments of Eire and the U.K. were also easily corrected since the displacement was a constant value. Italy was considered too deformed to be easily and accurately corrected so it was digitised, successfully, a second time when the digitiser fault had been rectified.

Once the individual countries had been checked, any straight line boundary segments temporarily inserted were removed, and GIMMS "basefiles" of country segments, each converted into node coordinate space, were created. These "basefiles", all in the same coordinate space, were stored in GIMMS internal file form and so could not be directly read or submitted to another program. A GIMMS routine for writing out its internal files into a more normal form was used so that a comprehensive set of EEC Level II coordinates, now all in the same coordinate space, could be stored and used instead of the original incompatible coordinate files. The GIMMS basefiles were concatenated and presented to GIMMS for polygon formation and the drawing of a Level II boundaries map of Europe. There were few problems at this stage. The GIMMS internal "polygon" file was stored ready for choropleth map generation.

The importance of studying regional variations at many different levels of aggregation has been described in Chapter 2. It would therefore be restrictive and unwise to preclude the capability of aggregating the Level II regions up to Level I and again to country level regions. Since the levels of the EEC are nested, one in another, and therefore make use of common boundary segments, it is theoretically a simple matter to aggregate upwards from the most disaggregated regional level, provided information on boundary definition is available. GIMMS allows such information to be incorporated within the segment labeling of the coordinate files so that each segment has a set of polygon labels associated with it - a right hand and a left hand label for each level in the hierarchy. The combination of right and left polygon labels for any one segment will be unique for the most disaggregated regional level, but will not necessarily be so for more aggregated levels. In fact, the right and left polygon labels for the latter will be identical if that segment does not form a boundary at that level.

Provided this hierarchical labelling has been incorporated, GIMMS will produce outline maps of the desired level of aggregation if supplied with information at its polygon building stage. This level selection facility is most effectively used by adopting a labelling principle which allows polygon inclusion or exclusion on the basis of alphabetical ordering. To acheive this for the regions of the EEC, the segments were labelled as demonstrated by the following example of the boundary between Scotland and England.

```
/CUSCOTLA/CUNORTHE/BU48/BU39/AU1/AU1/
|                  |         |       |
|   level II       | level I | state |
|                  |         | level |
```

By specifying, during the GIMMS polygon formation phase, the exclusion of all regions whose labels do not begin with the letter B, a map of level I region boundaries only will be created. The second letter of all polygon labels in the above example is U, indicating that all belong to the United Kingdom. The corresponding letter for French regions is F and for Germany, G and so on. This nomenclature also enables a single country to be selected and mapped from the data files, as opposed to the mapping of the entire EEC, by the same technique of alphabetic discrimination.

3.6 MANAGING THE DATA

1977 Social data for Levels I, II and member state level regions of the EEC was provided by the EUROSTAT publication "EUROSTAT 1977 Statistics: Regional Statistics on Population, Employment and Living Standards" (1979). Approximately 100 selected variables were recorded on punch-cards by the punching service at Durham Computer Unit, and then copied into disk files according to logical groupings of variables. The contents of the files were then manually checked with the EUROSTAT publication for punching errors. The variables selected include population statistics, employment/unemployment data, births and deaths, and age-structure data.

Within each file, the data took the form of a conventional case by variable matrix, the cases being the states, Level I and Level II regions of the EEC. An index was set up in a file for discovering which variables resided in which files. The order of the cases is important, since GIMMS expects to receive data for mapping in the same order as it produces the associated polygons. The order in which the polygons are produced may be controlled to some extent by requesting output in alphabetical order.

As regional levels are selected for choropleth maps, it was vital to ensure that the appropriate data values for the variable in each of these regions to be mapped are also selected. The ordering of cases within the data files was, therefore, also alphabetic, the same names being used as for GIMMS polygon labeling. Alphabetic sorting of these names resulted in a natural ordering, first by EEC Level, next by member state, and last by Level II region, as follows:

```
EEC - 9

                     ┌Belgium
State Level---------┤ ...
                     └U.K.


                     ┌Belgian regions
Level I-------------┤ ...
                     └U.K. regions


                     ┌Belgian regions
Level II------------┤ ...
                     └U.K. regions
```

The technique used for data selection simply made use of the operating system utility of defining portions of files by specifying line (or record) numbers. This means that explicit information about order and positioning of cases within each data file is required. In principle, however, a software solution to this problem is obvious: it would not be difficult to incorporate a program which selected data on a similar alphabetic discrimination process as GIMMS, or which simply referenced a look-up table to discover the line range of the relevant cases within a file. The same program could also include a facility for first discovering the appropriate data file in which the chosen variable resided, by using the existing index of variables.

Statistical procedures used were provided by MIDAS and CLUSTAN, which were fed with data on the same basis as GIMMS, that is via file name and line numbers; in principle, exactly the same proposed data selection techniques could be applied for these packages. MIDAS was chosen as a statistical package not only because it was designed for interactive use, but because it has the ability to select variable values from a number of data files and combine them into a single file which it produces as output. This facility was found to be particularly useful when combinations of variables were required for cluster analysis. MIDAS itself has clustering routines but they are by no means as comprehensive as those provided by CLUSTAN which is dedicated to this statistical task.

CLUSTAN is able to produce output in the form of "classification arrays" which are arrays of values indicating the cluster groups to which each individual belongs. For hierarchical techniques, these arrays will, of course vary according to the value of the criterion on which clustering was based. It is, therefore, necessary to define the number of groups required. A classification array from CLUSTAN, once re-formatted, may be submitted as data to GIMMS, and mapped in a similar way to any other variable, though, since the results are essentially numerical labels representing qualitatively different classes, the mapping symbolism should essentially be non-gray scale.

3.7 THE MAPPING OF SELECTED EEC REGIONAL VARIABLES

In order to evaluate the utility of an information system based upon the software discussed in this Chapter, eight variables from the 1977 data were selected for choropleth mapping. The variables chosen were as follows:

Population Density

Birth Rate

Mortality Rate

Unemployment Rate

Age Index

Employment in Agriculture

Employment in Industry

Gross Domestic Product

(See Appendix - B for variable definitions)

Some of the data for the above variables, for example unemployment figures, are derived from Community sample surveys, conducted in order to provide comparable data across the member states. Some of the EUROSTAT data were deliberately not selected for mapping because the figures are unsuitable for comparison. For example, some of the unemployment variables are based upon persons registered at labour exchanges; the values tend to reflect the degree of development of the labour exchange and are also governed by variations in administrative practices and laws from one country to the next.

Some of the variables used for mapping have missing values for certain regions. Where this was found to be the case, values were interpolated from Level I data. Details are given in Appendix - B.

The choropleth maps produced for the eight variables are displayed in Appendix - A. All of these are at the Level II degree of aggregation. Prior to submitting each variable to GIMMS for mapping, MIDAS was used to provide basic descriptive statistics and a histogram on which the selection of class intervals was based. The appropriate Level II region values were selected using the line-range facility of the MTS operating system. The first few maps were produced interactively, but this proved to be a time-wasting process, particularly when the computer was heavily used. The rest were processed in batch, using a standard GIMMS run set up in an MTS file and edited for the individual requirements of each map.

Output from GIMMS was in the form of an MTS plotfile. This is a common intermediate state produced by all supported graphics programs at NUMAC, which may be diverted to any selected graphical output device using the appropriate systems program to drive the device. Before sending these plotfiles to the hard-copy plotter, they were viewed on a Tektronix 4014 to check for any errors.

In addition to the eight maps of "simple" variables, two maps of "composite" variables, derived from the clustering of selected original variables, have also been produced. Composite variable "A" was formed from clustering of two variables: Unemployment Rate and Gross Domestic Product (GDP) per capita. Composite variable "B" was formed from clustering of five original variables: Unemployment Rate, GDP per capita, Age Index, Net Migration and Employment in Agriculture. The cluster analysis itself was performed by the package, CLUSTAN. The particular technique employed was Ward´s method, based on squared Euclidean distance. Selection of number of cluster groups to use for mapping was aided by reference to dendrograms produced by the package. The facility for output of classification arrays was made use of. A simple re-formatting program arranged the classification arrays in a form suitable for input to GIMMS.

None of the maps display shading for the Isle of Man. This is due to the lack of information about its inclusion or exclusion from the EEC regionalisation. No reference is made to the Isle of Man in the "EUROSTAT 1977 Statistics" publication.

The eight "simple variable" maps in Appendix-A generally show expected patterns of variation in respective variables over the EEC. The population density map demonstrates particularly the influence of size and placement of areal units. Berlin, Hamburg and Bremen appear as regions of higher population density than, say, the Paris region (Ile de France). This may be simply because the latter includes much of the surrounding rural environment. In contrast, Scotland appears to be very sparsely populated, the effects of the Edinburgh/Glasgow urban complex being counteracted by the rural northern area and islands.

The maps of birth rate, mortality rate and age index should, perhaps, be considered in association with one-another. Both southern Italy and Eire show high birth rate (perhaps the influence of Catholicism) supported by the high degree of "youthfulness" of population demonstrated by the age index map. Mortality rate is correspondingly low in these areas. The coastal region on the French/Italian border has low birth rate, high mortality rate and an ageing population. This is most probably due to its popularity as a retirement area. The same pattern exists to a lesser extent for the South-West region of the U.K.. Berlin and Hamburg also display this same pattern, but the explanation is very unlikely to be the same. Perhaps the inner areas of these cities house an older population, the younger people living in the surrounding regions. For Berlin, however, this is unlikely to be the case - it is more likely that young people have moved from the city altogether. Migration figures would show if this were true.

Employment in agriculture and emploment in industry show a distribution largely as expected, with heavy dependence on agriculture in rural regions of southern Italy, particularly inland, south and west France, and Eire, and a heavy dependence on industry in the North-West and West Midlands of the U.K., the regions around the Rhine and Ruhr valleys, and southern Germany. It is perhaps interesting to note that the city regions of Berlin, Hamburg and Bremen have values for industrial employment as low as those for Eire and Sardinia. This is presumably due to a high degree of employment in services in the former.

Unemployment rate has, perhaps, a less easily explained pattern of variation. High values tend to occur in agricultural regions, for example Eire and southern Italy, possibly due to the lack of alternative employment to agriculture. However, bearing in mind that the figures are expressed as a percentage of the total population, the reason for this pattern is more likely to be the high proportion of children in these areas. This is bourne out by the industrial regions of east Belgium having high unemployment values and also high birth rate values. The map demonstrates how secondary factors can strongly influence the pattern of variation of the primary variable. A more informative index would have been unemployment expressed as a percentage of working population. Values for Gross Domestic Product are high in prosperous urban regions of Berlin, Hamburg and Bremen, around Paris and around Copenhagen, and low in rural regions of southern Italy and Eire.

The symbolism selected for the two composite variable maps attempts to give the impression of seperation of regions into categories. The first of the two maps - Composite Varible A - was deliberataly used in order to compare the resulting regionalisation with a study documented in an EEC report - "The Regions of Europe" (Jan 1981). This study maps a composite indicator formed from GDP per capita and long-term unemployment rate. No explicit definitions of the two variables are given. The composite indicator used is an interval scale variable and not categorical as in this study. The regionalisations apparent from the EEC study and this study do show similarities, for example, Germany and mid-France are classified together in both maps, as are southern Italy and Eire. South-west France also forms a distinct region.

Composite Variable - B, shows a fairly similar pattern, with southern Italy and Eire remaining together in the same group and a larger area of south and west France appearing as a distinct region, presumably due to the influence of employment in agriculture. Germany is a little more broken up, but much of the country forms a region with north-east France.

It was not the primary concern of this study to analyse the patterns shown by the variables, nor their implications, but to concentrate upon the techniques required to present the data. However, the brief discussion included above serves to demonstrate the importance of viewing the spatial patterns in context, and also to point out some of the common pitfalls in the interpretation of such patterns.

CHAPTER 4

DISCUSSION AND CONCLUSION

4.1 AN EVALUATION OF THE SOFTWARE, AND SOME SUGGESTIONS FOR ITS IMPLEMENTATION

The practical work undertaken in this study has attempted to demonstrate that there exists a suitable resource base in terms of computer software on which to build an efficient geographical information system for the EEC. Given that the data provision organisations continue to strive for comparibility and consistancy by conforming to standards set up by the central data-gathering authority, EUROSTAT, the suggested system incorporating the packages MIDAS, GIMMS and CLUSTAN should be able to provide a means for performing fairly detailed research into EEC regional data. However, it remains to discuss what disparity, if any, exists between the proposed software system and an ideal one.

The MTS operating system provides very suitable interactive facilities for the purpose. Programs can be implemented easily which interrogate the user and call up the necessary routines or data files appropriate to the answers supplied. Given that MTS has been written for general-purpose use, there are very few drawbacks in its adoption as a basis for a geographical information system, (see section 3.2). Depending upon the rationale adopted for data organisation, there might be more appropriate operating systems, for example, utilising the hierarchical file structures of the UNIX operating system. However, a more important issue is probably to design the information system so that it can use as many common operating systems as possible; MTS is considered a fair representative, there being a number of operating systems which provide similar facilities, for example, the DEC-system "TOPS-20". It should, perhaps, be mentioned that the general unsuitability of most computers for data-base management, particularly the relational type, has been recognised, and there has been recent interest in the design of "data-base machines", the architecture of which is more suitable for that purpose, (Connett, 1981).

Interactive statistics packages are increasing in availability. An interactive version of SPSS(Statistical Package for Social Scientists), called SCSS (see Nie et al, 1980), and others such as MINITAB and P-STAT are relatively well-known ones. Most of these packages have the advantage of being portable, a quality which is not a feature of MIDAS, written specifically for implementation at MTS sites. MIDAS was used in this study in preference to SPSS, firstly because it is interactive and secondly because of some useful data management capabilities, (see section 3.6). Many of the features of MIDAS are paralleled, or should be in the near future, by more portable interactive statistics packages which should be easily implemented on machines used for research within the EEC. CLUSTAN provides a substantial collection of clustering routines which should be sufficient for most purposes. It is not, however, designed to be interactive, although it is portable and generally available for batch use.

The importance of the exploratory approach to data analysis has been mentioned (section 1.2.1). Connected with this is the unsuitability of some data for undergoing commonly used parametric and even non-parametric tests. Social science data is commonly right-skewed, for example, census variables such as number of persons per room or number of children per household, and spatial data tends by nature to be autocorrelated, so that important assumptions of conventional statistical techniques are hard to meet. For this reason, the use of more robust statistical measures are being encouraged, (Erickson and Nosanchuk, 1979). It would seem relevant to incorporate such techniques into a geographical information system for spatial socio-economic data. However, I am as yet unaware of a generally available package which offers such facilities.

The state in which the coordinate data is made available to users of the information system requires some discussion. For map production using the EUROSTAT Levels, there seems little point in providing the user with direct access to the raw coordinate data, since he would have to spend time waiting for GIMMS to create first a "basefile" and then a "polygon" file, which is not inconsiderable given the quantity of coordinate data involved. There might be some case for the user creating his own "polygon" file, since it is at this stage that the selection of EUROSTAT Level or member states to be mapped occurs, although this could be overcome by creating all possible combinations of polygon files and allowing the user to chose the one required. The decision seems to reduce to pay-offs between flexibility for the user, time spent by the user in map production, and storage space taken up by the information system. Perhaps the best rationale to adopt would be to provide the standared GIMMS "polygon" files for Levels II, I, and state level for the entire EEC, assuming that these will be in highest demand, but to allow access to the "basefile" should the user require to map selected countries. Access to the raw coordinate data should perhaps be made possible for those users who need to carry out more specific tasks outside using the standard EUROSTAT regions. Even when provided with a GIMMS polygon file, the map-creating phase itself can be a little intricate for the first-time user. GIMMS provides certain defaults for interval selection, shading type and character sets. It must be mentioned, however, that the default interval selection must be used with care; misleading maps can result if these are not sensibly determined, (Evans, 1976). It would be an advantage if GIMMS incorporated a closer link between histogram production and the selection of class intervals by automatically generating intervals based upon standard deviations, geometric progressions and percentiles at the request of the user. A facility for drawing the histogram on the final map would be beneficial. Also some means of coping with flow data, such as the drawing of flow arrows for migration between regions, might be a desirable feature for future versions of GIMMS.

There has been some discussion in previous sections of this study (1.2.2) about the provision of facilities for unsophisicated computer users. Although MTS itself has been written as a "user-friendly" operating system, it is still necessary, for the purposes of the EEC information system, to screen the user even further from the data retrieval and program-linking procedures so that he may, for example, ask for a certain variable by name without needing to know which file it resides in. The interface written need not necessarily overide the command languages of GIMMS and MIDAS since these are designed to be simple to use. The only conceivable improvement to MIDAS might be to make it query/response-driven rather than command-driven. The GIMMS-3 command language could be made more helpful; unless the user is very familiar with the command syntax, the manual must be kept at hand for reference. CLUSTAN is also command-driven, but the commands are not free-format or easy to understand without the aid of a manual. Requirements are specified by placing numbers or letters in specific columns of the command line, so that constant reference to the manual is necessary even for the most familiar user. The reason may be because of the sheer amount of computation required to perform the clustering tasks.

Apart from some problems with CLUSTAN, a workable, if not ideal, user interface may be created encompassing initial introduction to the information system, data retrieval by variable name, and geographic individual requirement by name, for example, "Level I" or "Level II". Once variable and region level had been selected, a user could be linked directly into the GIMMS command language for map production, the relevant "polygon" file and data file being submitted automatically to the package by the interface program. The same procedure could be adopted for linking to the statistics package, submitting the selected data automatically to MIDAS when "statistical analysis" has been requested by the user. The obvious problem with this kind of "labour/saving" interface is that the user has to be familiar with more than one interface language, which could be confusing. It would be an advantage if an interfacing standard could be formulated for packages and data management systems. At present, data destined for input to CLUSTAN and the associated commands would have to be sent to a batch queue, or incorporated into a file which could then be "executed" by the operating system at the user´s terminal. A interactive pre-processor interface could be written for CLUSTAN, which set up such a file of commands based upon user responses to a sequence of questions. It would be helpful to allow batch use of the information system, or selected sections of it, since such procedures as map production can take up to an hour of terminal time to complete on a heavily used system. Finally, since the information system should be available for all member states of the EEC, it is important to provide interfacing in at least three languages, the user making the selection by responding to an initial question given in all the language alternatives.

4.2 MAKING THE SYSTEM AVAILABLE

Once an EEC information system for regional socio-economic data has been designed and written, and exists as a portable, available product, some consideration must be given to the practical implications of making it available to users. The continued data gathering and data provision role of EUROSTAT is really fundamental to the entire scheme. So also is the future provision of data in computer-compatible form. EUROSTAT will hopefully continue to take responsibilty for regular updating and maintainance of confidentiality of information.

The EEC has a Central Computer Center which houses an ICL 2980 machine. EUROSTAT has created a data base for economic statistics such as trade figures, consumer prices, and balance of payments, called CRONOS which allows information retrieval to be requested in English, French and German. This system runs only at the Computer Center and, up until recently, has not been made generally available to external users. The EURONET packet-switching network aims to connect any EURONET-compatible terminal with a number of host machines, notably the ICL2980 running CRONOS, so that users anywhere in the EEC may obtain direct access to up-to-date information provided by EUROSTAT. the operation of the telecommunications network has been undertaken by the post and telephone authorities of the nine member states. (See Mesnage, 1980)

The advantages of such a wide-reaching network are considerable, enabling fast and efficient dissemination of information, especially important for those requiring immediate details of up-to-date statistics which may vary from week to week. The cost of installing the EURONET network has been bourne by the European Commission, but the cost of linking into a node by telephone, particularly for the far corners of the EEC, may not be inconsiderable. This may discourage academic and other, perhaps less affluent, institutions from using EURONET. Depending upon the cost, therefore, such a network system may not be the ideal mechanism for providing European socio-economic data, along with the information system envisaged, to those interested parties. In addition, it is unlikely that many socio-economic regional statistics, such as population counts, average familiy income, or housing data, particularly at the more disaggregated levels, are capable of being updated more frequently than once per year, so that there is not so much urgency for frequent and direct reference to EUROSTAT as the data gathering agency. It is unlikely, too, that regional planning decisions, for example, would be made on the basis of weekly or even monthly fluctuations in social statistics. In short, it may be more cost-effective to supply the information system to computer centers throughout the EEC as a standard, portable, "parcel", providing automatic updates of statistics via magnetic tape on a regular basis. The system would obviously have to be easy to implement, and also easily extendable so that a time-series data base of regional socio-economic statistics could be built up at each site as information updates were received. The system should also allow data output onto transportable magnetic media in some standard format so that EUROSTAT themselves could receive the regular and efficient feedback of local data which they require from each member country. It may also be beneficial if each site were able to "attach" their own program routines or data to the information system and "mold it" to their specific requirements.

4.3 SUMMARY AND CONCLUSION

This study set out to discover whether there exists suitable software upon which to base a geographical information system for EEC regional socio-economic data, and to evaluate the feasibility of setting one up. Selected software was tested for suitability by requireing it to perform some chosen tasks which researchers and EEC decision-makers might wish to undertake. For this purpose, the EEC Level II region boundaries were digitised and associated attribute data organised in MTS line files. By making use of the MTS operating system facilities and linking together three major packages, GIMMS, MIDAS and CLUSTAN, via simple Fortran programs, ten maps displaying regional variations in selected simple and composite variables were produced.

The advantages and disadvantages of the three packages and the operating system have been discussed individually in section 4.1. Together, they provide fairly adequate facilities for regional research, allowing data manipulation, conventional statistical analysis techniques and mapping of original and "secondary" variables to be carried out with relative ease, given that no all-encompassing user interface has been written. The major difficulties experienced were in "once-off" data preparation, data cleaning and data organisation tasks such as boundary digitising and coordinate input to GIMMS. Organising attribute data in MTS files was much more time-consuming and tedious than difficult, the sort/merge facility being capable of coping with more complicated situations such as multiple records per case. Facilities missing from the software include such tasks as polygon overlay and point-in-polygon routines which may be required if point data or data pertaining to different regionalisations were incorporated. Also missing are routines to perform data analysis using more robust statistical measures, and facilities for determining extent of spatial autocorrelation, for example. However, the initial groundwork and important major features are provided for; more specialised routines can be incorporated at a later stage, provided that the system will be easily extendable. The remaining hard work lies in the design and provision of a user interface, whether it is simply to provide links between packages or whether it is to overide those of the packages as well. In addition, some efficient design for attribute data organisation is essential as the data base grows to include time-variable as well as space-variable data, so that the user may have quick and easy access to whatever portion of the data he requires. There should also be provision within the interfacing program to alert the user to incidence and extent of missing data or data which are estimates based upon survey samples.

This study has demonstrated that, by using the facilities of ready-written software, a much-needed geographical information system for EEC regional socio-economic data could be created with relatively little extra work. The utility of such a system seems considerable. Provision of up-to-date, easily accessible data sets along with associated facilities for manipulating, analysing and mapping these, is surely essential in order to permit efficient and timely regional policy-making by EEC planning authorities. Such a system would also be of considerable importance to the research bodies of Europe, much of whose time is taken up with data preparation and organisation tasks. The centralisation of such activities leaves much more time for actual research, and consequently increases the efficiency of any information feedback which they might provide.

References

CONNETT, J. (1981) "Data Base Software or Data Base Machine?" Paper delivered at the IUCC Conference, September.

ERICKSON, B.H. and NOSANCHUK, T.A. (1979) "Understanding Data". Open University Press.

EVANS, I.S. (1976) "The Selection of Class Intervals". Trans. Institute of British Geographers, volume 2(2), p98-123.

MESNAGE, M. (1980) "Dissemination of Statistics using Computer Facilities in the EUROSTAT". European Political Data, number 35, p17-21.

NIE, N.H. et al (1980) "A User's Guide to the SCSS Conversational System".

BIBLIOGRAPHY

BAXTER, R.S. (1976) "Statistical and Computer Techniques for Planners". Methuen.

CLIFF, A.D. and ORD, J.K. (1975) "Model Building and the Analysis of Spatial Patterns in Human Geography". Journal of the Royal Statistical Society, series B, volume 37, p297-347.

CONNETT, J. (1981) "Data Base Software or Data Base Machine?" Paper delivered at the IUCC Conference, September.

COOK, B. (1979) "Land Use on the South Coast of New South Wales" Chapter 5.

CORBETT, J. and FARNSWORTH, G. (1979) "Theoretical Basis of DIME in U.S. Bureau of the Census". In U.S. Department of Commerce Conference Proceedings. International DIME Colloquium, Washington D.C..

COX N. (1978) "Exploratory Data Analysis for Geographers". Journal of Geography in Higher Education. Volume 2(2) p51-54.

DAVIS, G. (1979) "The EURONET Story". European Political Data, number 31.

DAVIS, J.C. and McCOLOUGH, M.J. (1975) ed. "Display and Analysis of Spatial Data". Wiley, London.

DUEKER, K.J. (1979) "Land Resource Information Systems: A Review of Fifteen Years Experience". Geo-Processing, 1, p105-128.

ELSEY, T. GLYNN, S. and MILNE W. (1980) "An Algorithms Library for Spatial Data" IUCC Bulletin Volume 2(3).

ERICKSON, B.H. and NOSANCHUK, T.A. (1979) "Understanding Data". Open University Press.

EVANS, I.S. (1976) "The Selection of Class Intervals". Trans. Institute of British Geographers, volume 2(2), p98-123.

FOX, D.J. (1978) "MIDAS Reference Manual". Statistical Research Laboratory, University of Michigan, Ann Arbor, Mich..

GO, A. , STONEBRAKER, M. and WILLIAMS, C. (1975) "An Approach to Implementing a Geo-data System". Presented at the _16th IFIP/IAG Data Base Workshop_.

HANDLEY, D. (1976) "A File on the Regions of 5 EEC Countries". _European Political Data_, number 18.

HANDLEY, D. (1975) "Second Survey of Computing Resources in the ECPR Membership 1974". _Eurpoean Political Data_, number 14.

HARRISON J.C. (1979) "The LAMIS System". _Chartered Land Surveyor/Mineral Surveyor_. Volume 1(2)

HAY, D. (1977) "European Data Sources and their Limitations for Comparative Research". Paper presented at the Bergen Conference on _Regional Data Bases and Automated Cartography_.

LA COUR, A. (1980) "Policy and Recent Developments in EUROSTAT - Methods of Disseminating Statistics". _European Political Data_, number 35, p7-16.

LORENTZEN, J. and ROKKAN, S. (1976) "A Multi-level Data Base for Computer Mapping of Regional Variations in Western Europe". _European Political Data_, number 19.

MARBLE, D. and PEUQUET, D. (1978) "Problems in the Storage and Manipulation of Large Spatial Data Sets". Proceedings of the _UNESCO Conference on Computer Mapping for Resource Analysis_.

MARTIN, J. (1975) "Computer Data Base Organisation". Prentice-Hall.

MESNAGE, M. (1980) "Dissemination of Statistics using Computer Facilities in the EUROSTAT". _European Political Data_, number 35.

MOCHMANN, E. and MULLER, P.J. (1978) "Emerging Data Protection and the Social Sciences Need for Access to Data". Report on _IFDO Conference_, Cologne, 1978.

MONKHOUSE, F.J. and WILKINSON, H.R. (1971) "Maps and Diagrams". Methuen.

NAG, and WAGLE, S. (1979) "Geographic Data Processing". _Computing_

Surveys, 11, p139-181.

NEWMAN, W.M. and SPROULL, R.F. (1980) "Principles of Interactive Computer Graphics", 2nd edition, McGraw-Hill.

NIE, N.H. et al (1980) "A User´s Guide to the SCSS Conversational System".

NORRIS, P, RHIND, D.W. and HUDSON, R. (1980) "Alchemy in Action? Regional Data Bases, Statistical Analysis and Cartography in a Computer Environment". Paper delivered at the IFDO Conference, Turin, March.

OPENSHAW, S. (1977) "A Geographical Solution to Scale and Aggregation Problems in Region-building, Partitioning and Spatial Modelling". Trans; Institute of British Geographers. New series, p459-472.

PEUKER, T.K. and CHRISMAN, N. (1975) "Cartographic Data Structures". American Cartographer. Volume 2, p55-69.

PEUQUET, D. (1979) "Raster Processing: An Alternative Approach to Implementing a Geo-data System". American Cartographer. Volume 6(2), p129-139.

PITKIN, W. (1981) "Panda - A Data Base Query Language for Ordinary People" Paper presented at the IUCC Conference.

REVE, T. (1975) "Computer Mapping for Spatially Variable Data" European Political Data, number 17.

RHIND, D.W. (1980) "A Multi-temporal, multi-purpose, multi-user, multi-resolution geographical information system". In URPIS 7 - Proc. Australian Urban and Regional Information Systems Association Annual Conference, Newcastle, NSW.

RHIND, D.W. (1977) "Computer-Aided Cartography". Trans; Institute of British Geographers. New Series, volume 2(1), p71-97.

RHIND, D.W. (1978) "Why Digitise?". Area, 11, 3, p211-213.

RHIND, D.W. (1976) "Towards Universal Intelligent and Usable Automated Cartographic Systems". ITC Journal. Number 5, p515-545.

ROBINSON, A. and PETCHENIK, B. (1975) "The Map as a Communication System". Cartographic Journal, June, p7-14.

ROBINSON, A., SALE, R.D., and MORRISON, J.L. (1979) "Elements of Cartography", Wiley, 4th edition.

ROKKAN, S. and WAGTSKJOLD, J. (1979) "Joint Nordic Data Base for Regional Time Series" European Political Data, number 30.

ROKKAN, S. (1980) "IFDO Symposium on the Development of Joint Data Bases for Regional Analysis and Computer Cartography". European Political Data, number 35

ROKKAN, S. (1977) "The Council of Europe Regionalisation Scheme". Paper presented at the Bergen Conference on Regional Data Bases and Automated Cartography.

SANDE, T. (1977) "Cartographic Data Base for Time-variable Data". European Political Data, number 25.

SANDE, T. and ROKKAN, S. (1980) European Political Data, number 34.

SCHMITT, A.H. and ZAFFT, W.A. (1975) "Programs of the Harvard University Laboratory for Computer Graphics and Spatial Analysis". In "Display and Analysis of Spatial Data" by Davis, J.C. and McCollagh, M.J. (eds.), Wiley.

SPICER, J.I. et al (1979) "Information Systems for Government". HMSO.

TANENBAUM, E. and TAYLOR, M. (1980) Paper delivered at the IFDO Conference, Turin, March.

THOMPSON, C.N. (1979) "The Ordnance Survey Topographic Data Base: Concepts for the 1980´s". Paper presented at the 2nd United Nations Regional Conference for the Americas, Mexico City.

TOBLER, W. (1973) "Choropleth Maps Without Class Intervals?" Geogr. Anal., 5, 3, p262

TOMLINSON, R.F, CALKINS, H.W. and MARBLE, D.F. (1976) "Computer Handling of Geographic Data". UNESCO Press.

TUKEY, W. (1977) "Exploratory Data Analysis". Addison-Wesley.

VISVALINGHAM, M. (1978) "The Signed Chi-squared Measure for Mapping" Cartographic Journal. Volume 15(2), p93-98.

VISVALINGHAM, M. and DEWDNEY, J.C. (1978) "The effects of the size of Areal Units on Ratio and Chi-square Mapping". Census Research Unit Working Paper, number 10, University of Durham.

WALLACE, M. (1981) "Natural Language Access to a Relational Data Base". Paper presented at the IUCC Conference.

WAUGH, T.C. (1978) "GIMMS Reference Manual" Inter University/Research Council Series Report 30, Program Library Unit, University of Edinburgh.

WEBSTER, R. (1977) "Quantitative and Numerical Methods in Soil Classification and Survey". Oxford: Clarendon Press.

WISHART, D. (1978) "CLUSTAN User Manual", 3rd edition. Inter University/Research Council Series Report 47

YULE, G.U. and KENDALL, M.G. (1950) "An Introduction to the Theory of Statistics". Charles Griffin.

EEC Report "The Regions of Europe". (January, 1981) p113-118

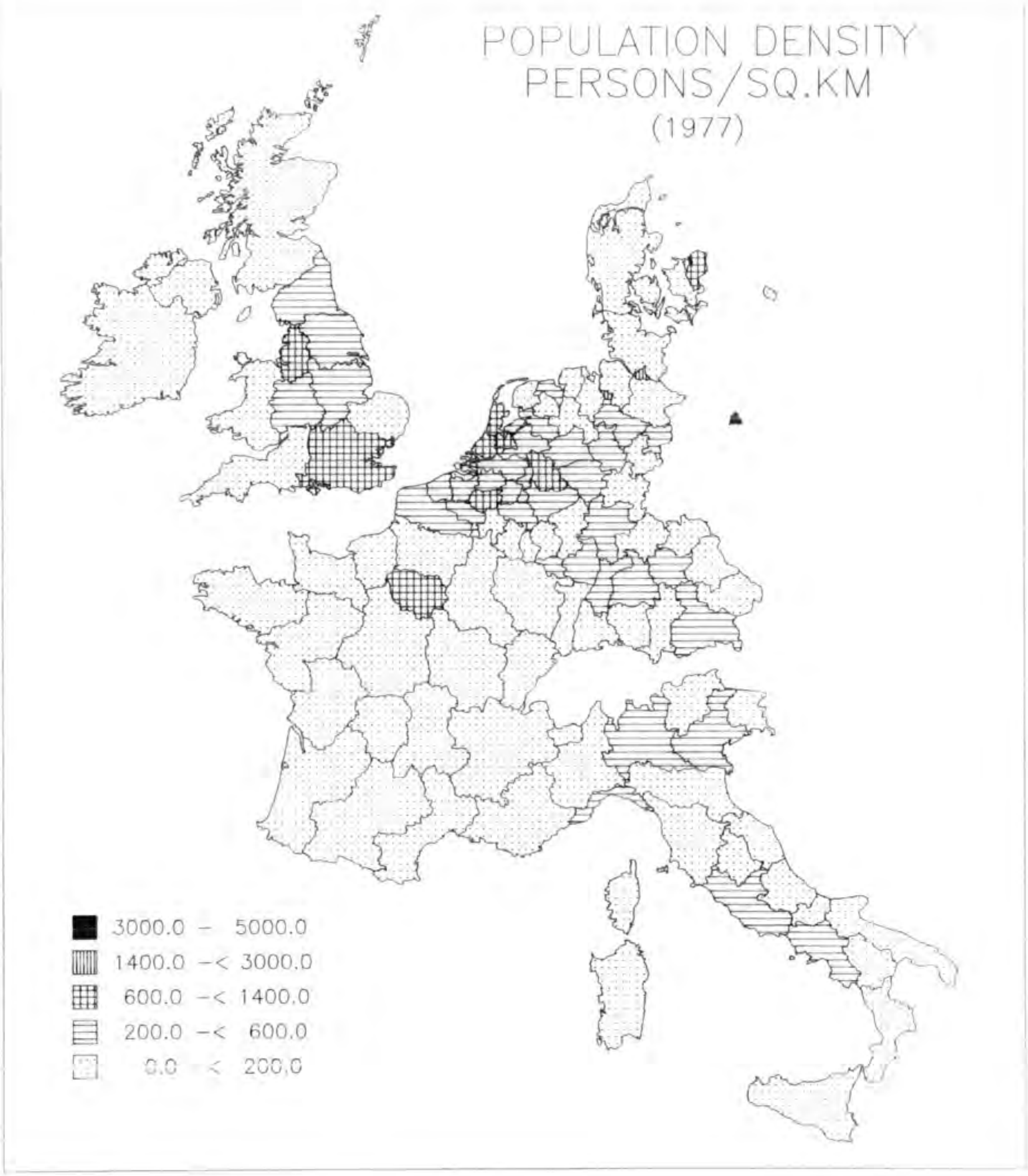
Laboratory for Computer Graphics and Spatial Analysis (1971) "LAB-LOG". Harvard University, Cambridge, Mass..

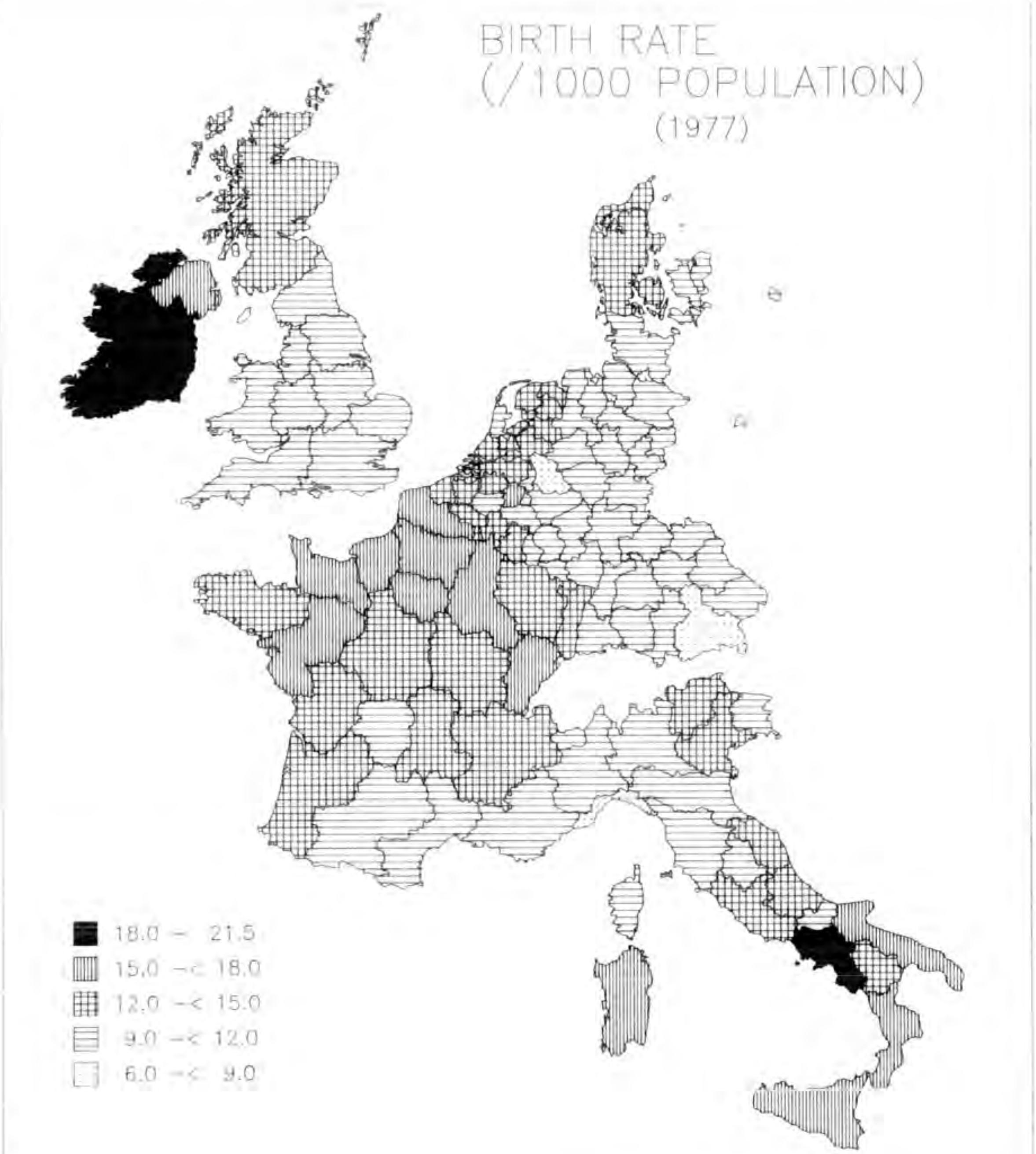
Report on the "Meeting of Committee of European Social Science Data Archives". European Political Data, number 22 (1977)

Report on the "European Working Group on Data Bases for Regional Analysis and the "European Meeting on Regional Data Bases for Automated Cartography" European Political Data, number 25 (1977)

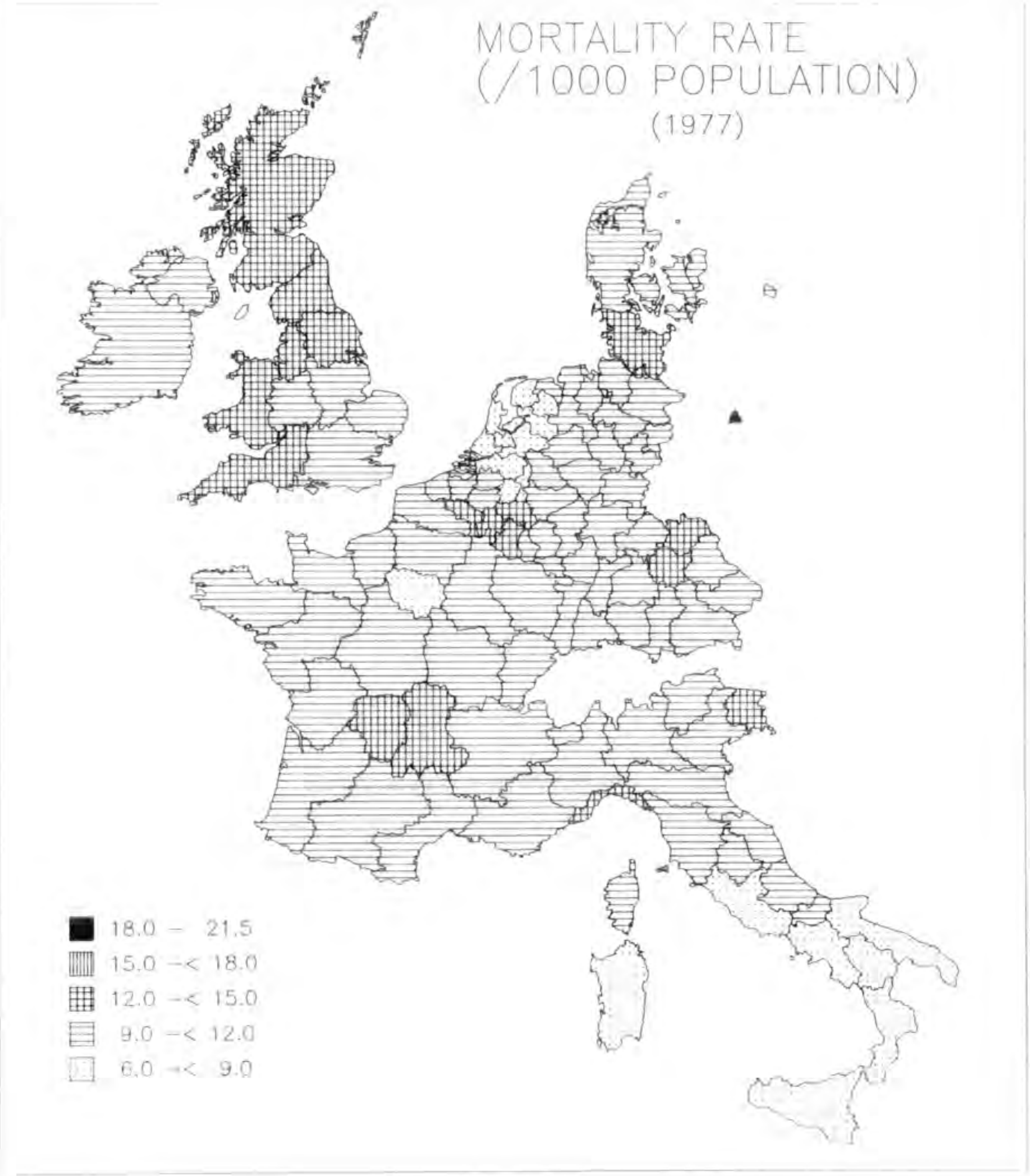
Report on "Data Bases for the Regions of Europe". European Political Data, number 27 (1978)

APPENDIX - A

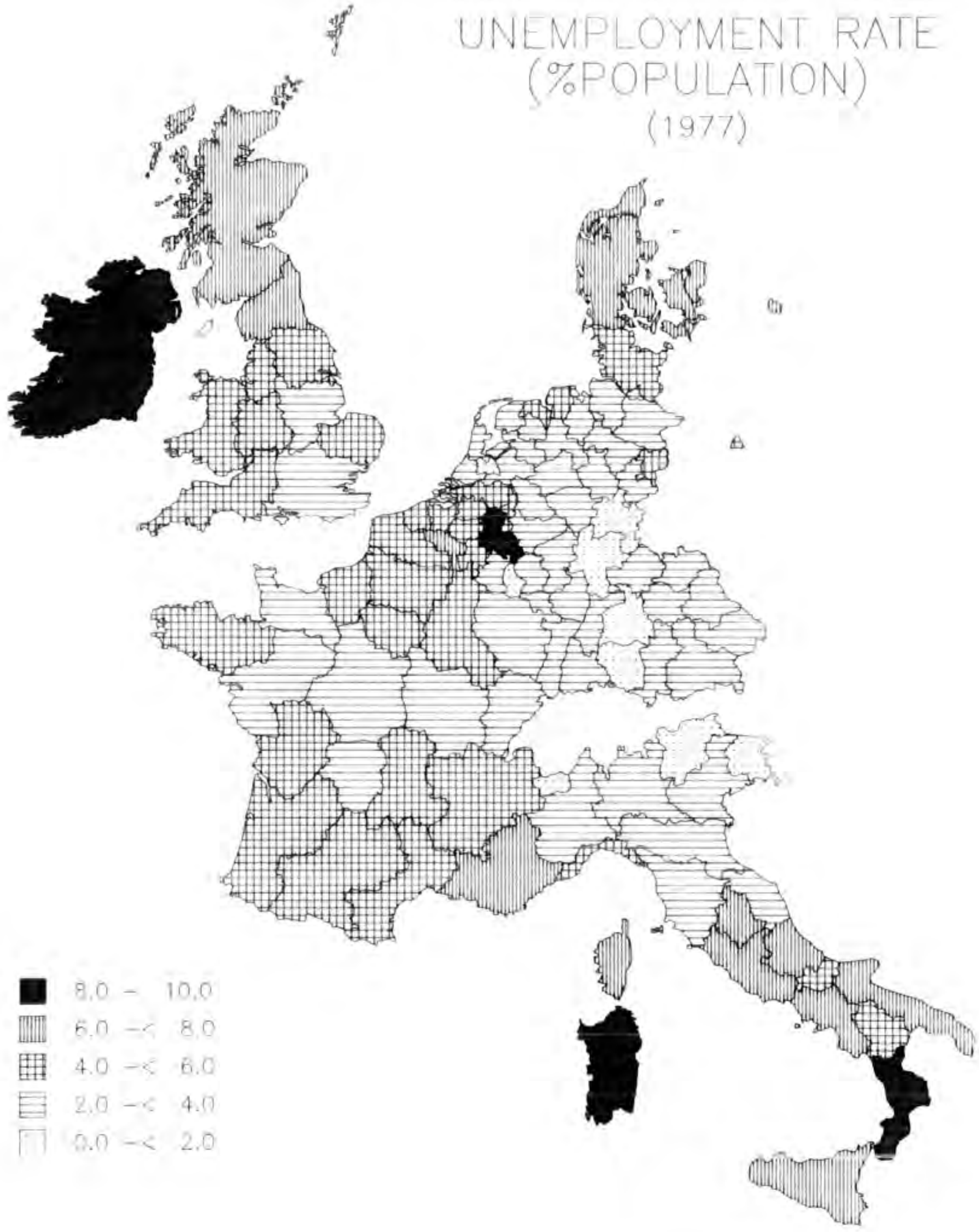

POPULATION DENSITY
PERSONS/SQ.KM
(1977)
3000.0 – 5000.0
1400.0 –< 3000.0
600.0 –< 1400.0
200.0 –< 600.0
0.0 –< 200.0

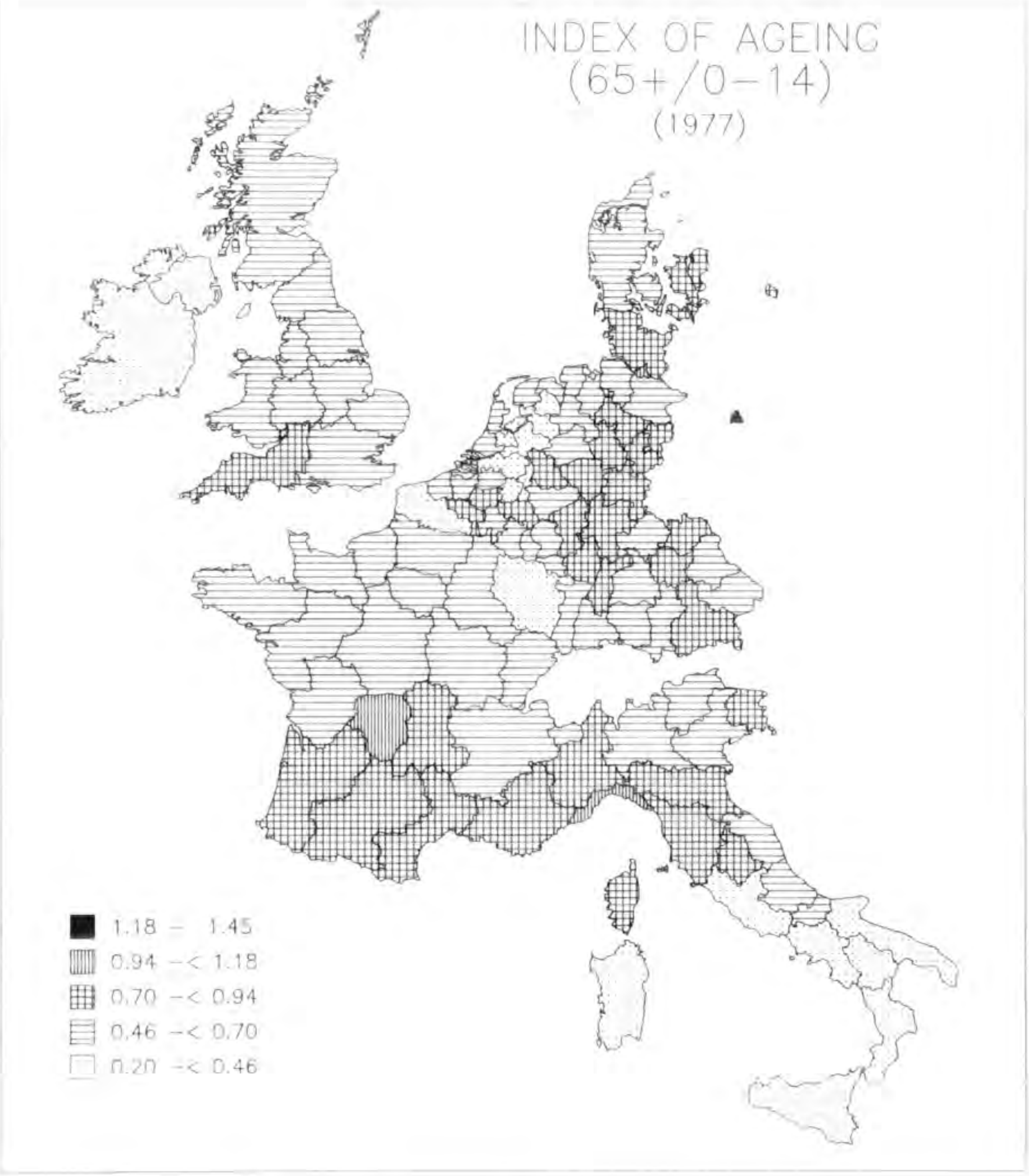


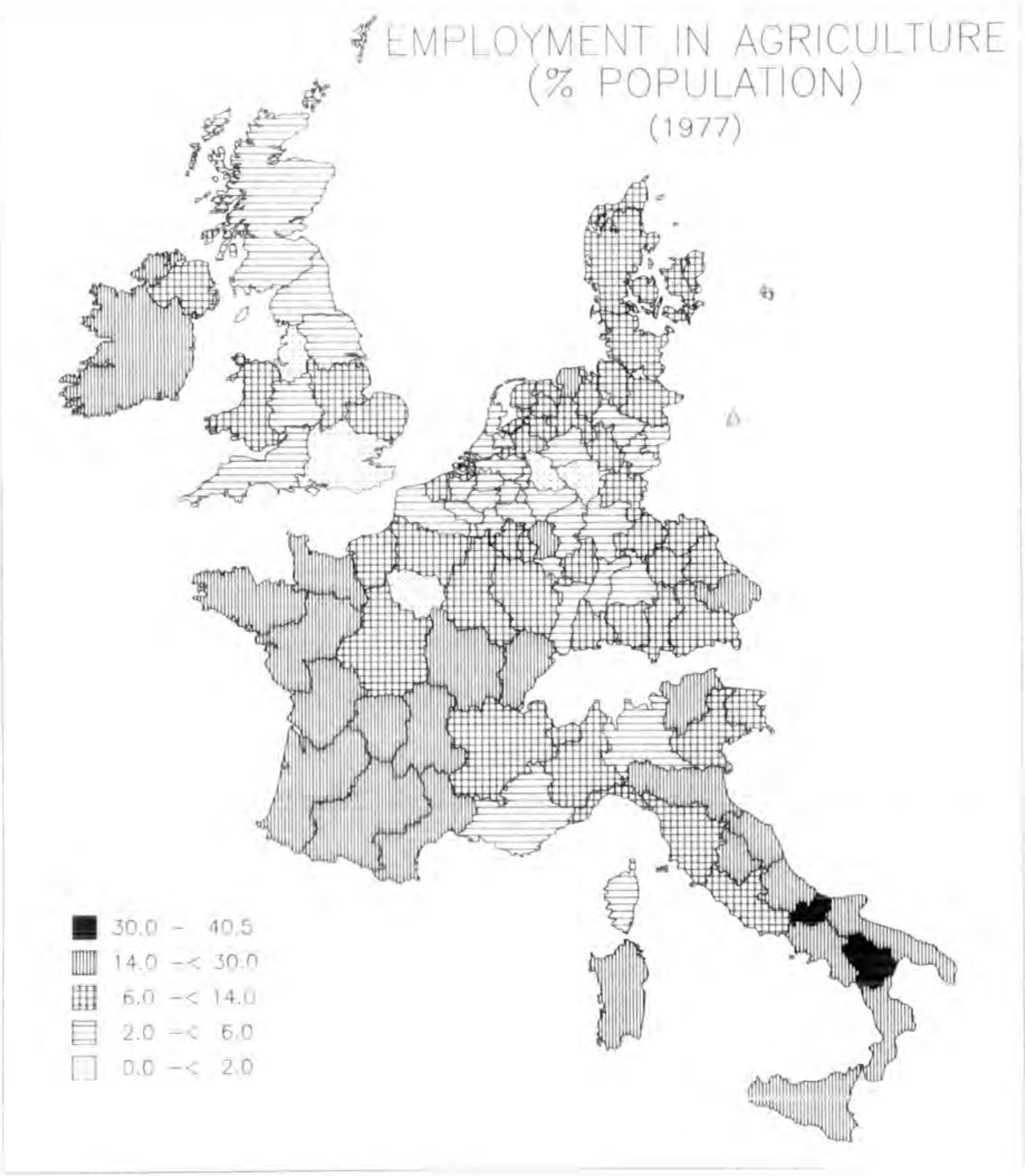
BIRTH RATE
(/1000 POPULATION)
(1977)


MORTALITY RATE
(/1000 POPULATION)
(1977)
■ 18.0 – 21.5
▥ 15.0 –< 18.0
▦ 12.0 –< 15.0
▤ 9.0 –< 12.0
▒ 6.0 –< 9.0

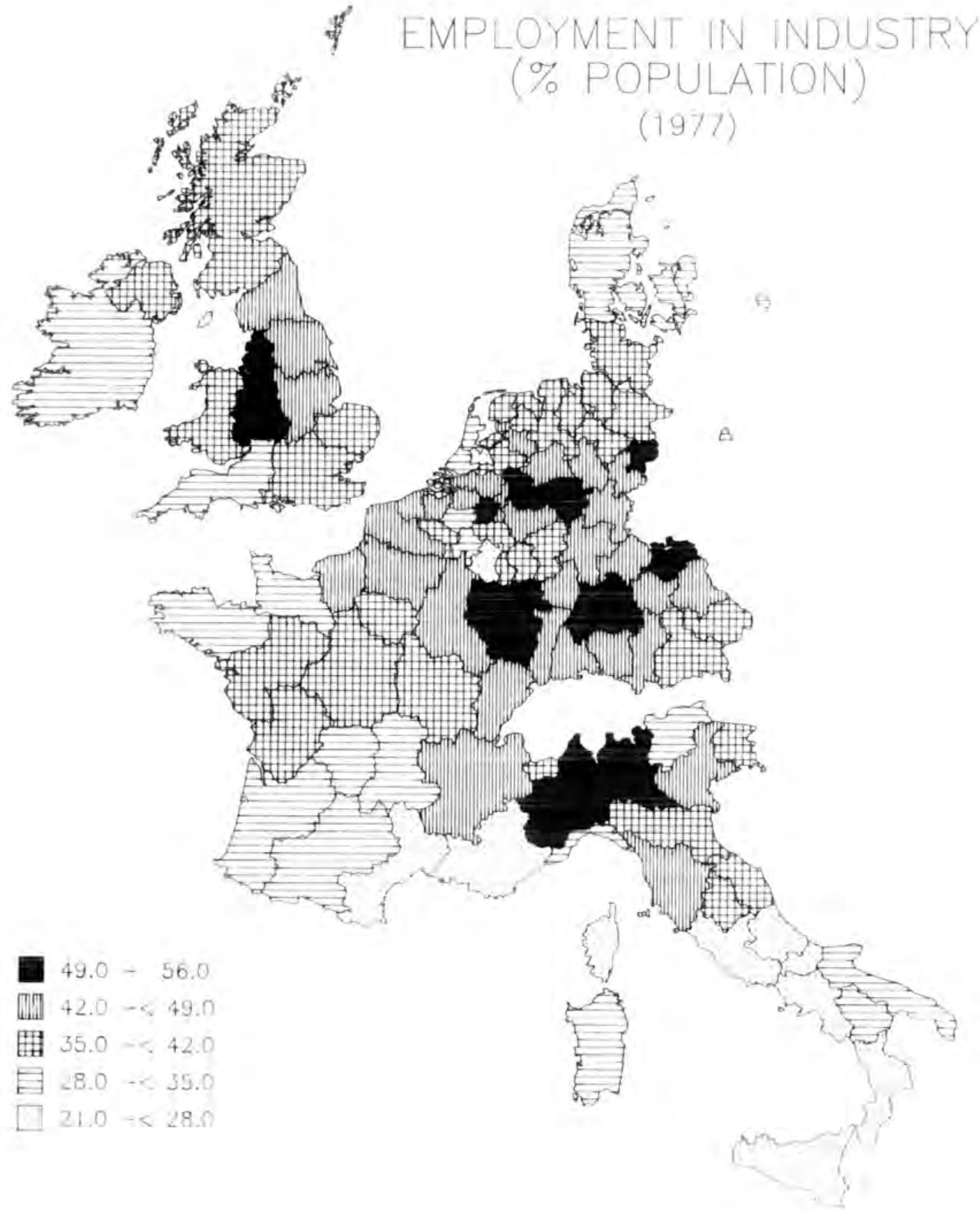
UNEMPLOYMENT RATE (%POPULATION) (1977)



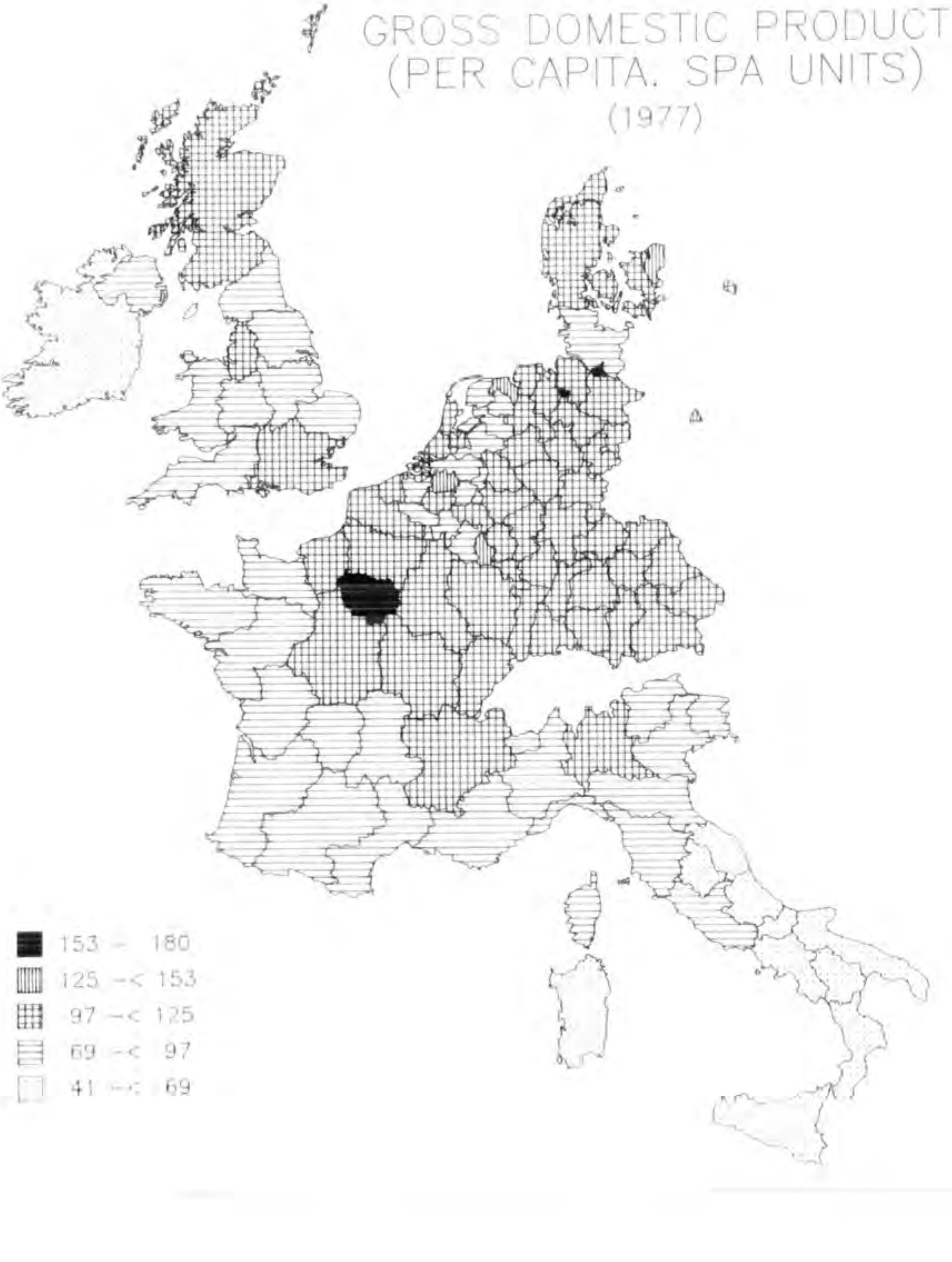

INDEX OF AGEING
(65+/0–14)
(1977)
■ 1.18 – 1.45
▥ 0.94 –< 1.18
▦ 0.70 –< 0.94
▤ 0.46 –< 0.70
□ 0.20 –< 0.46



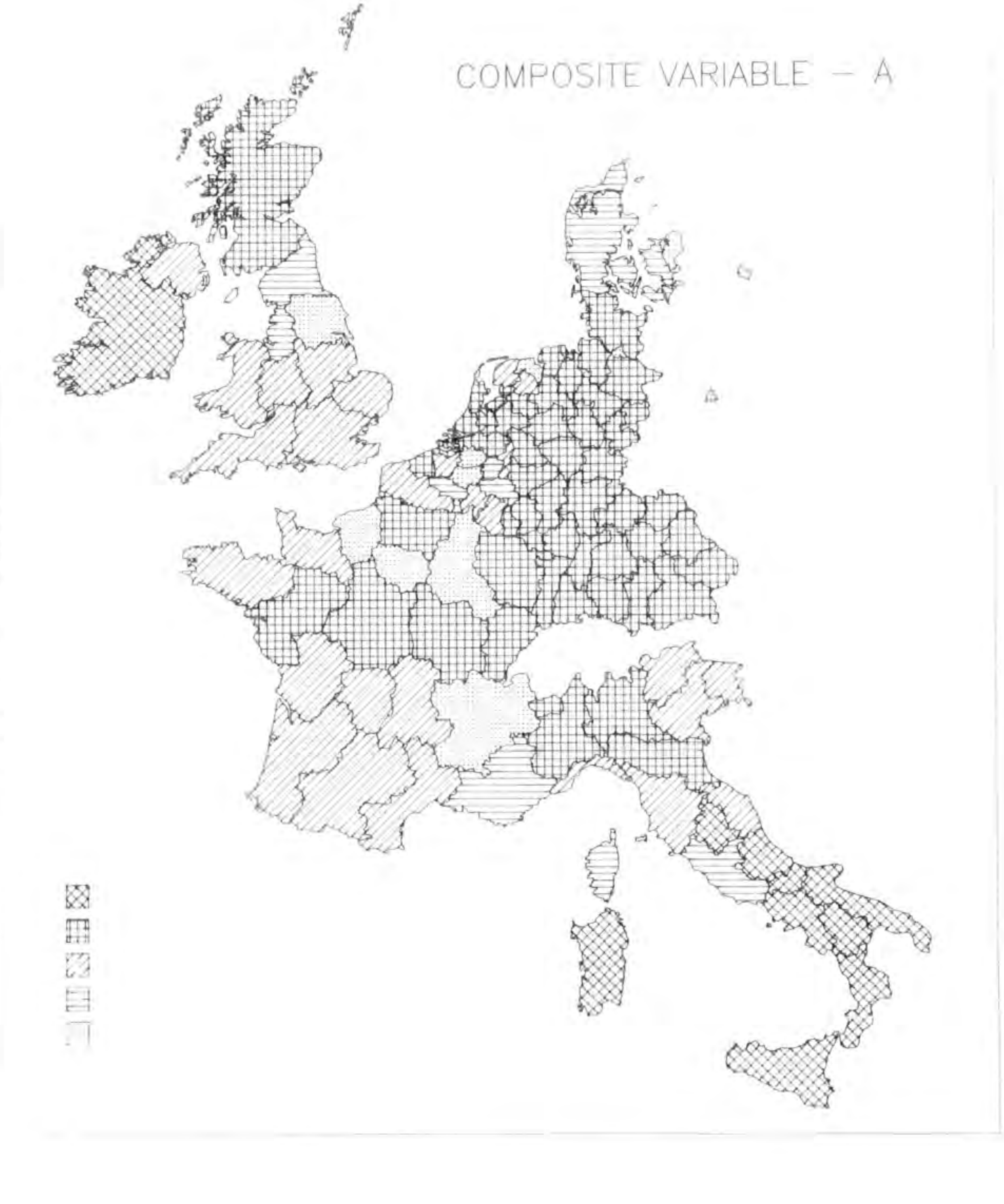
EMPLOYMENT IN AGRICULTURE (% POPULATION) (1977)



EMPLOYMENT IN INDUSTRY (% POPULATION) (1977)

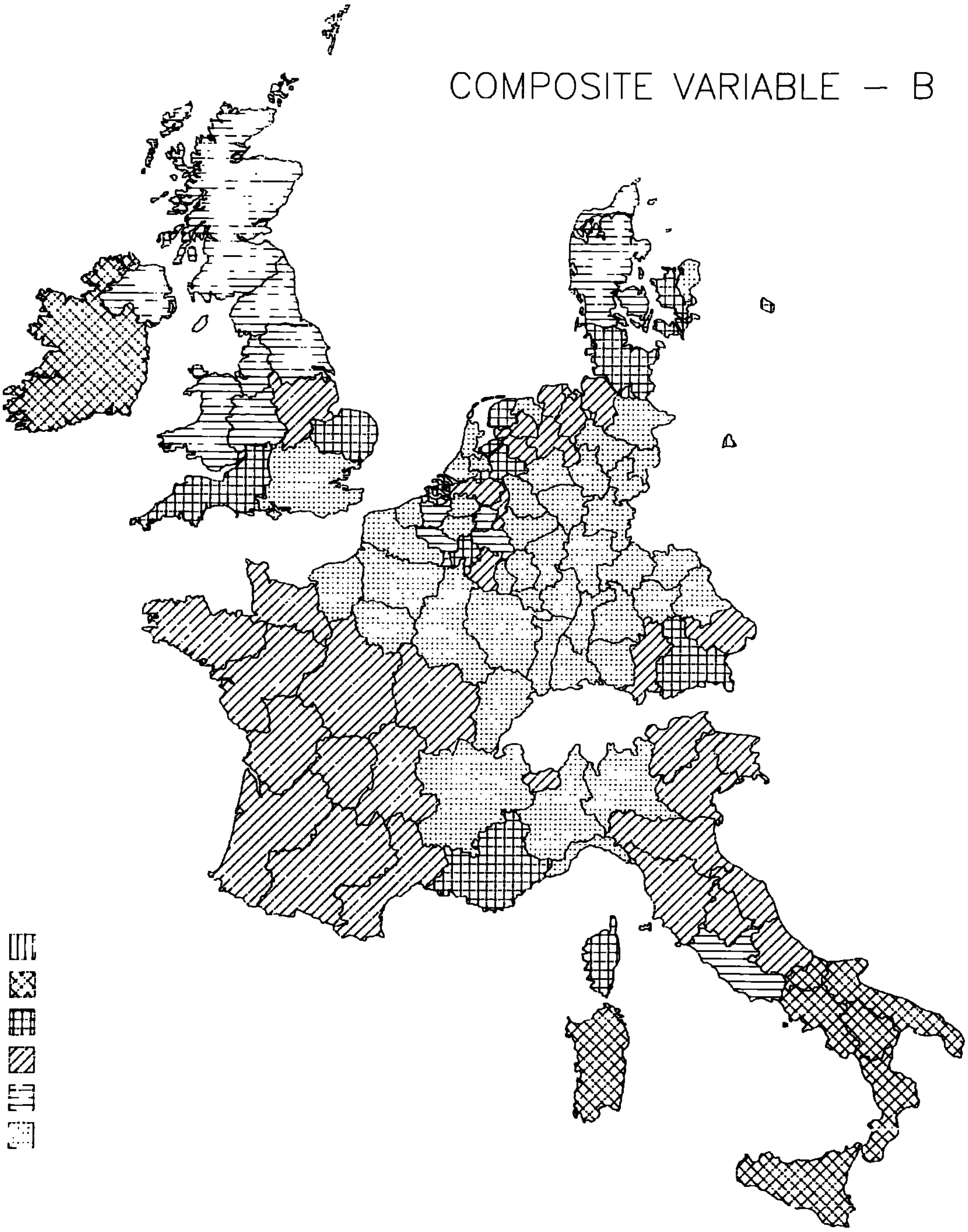


GROSS DOMESTIC PRODUCT
(PER CAPITA. SPA UNITS)
(1977)



COMPOSITE VARIABLE — A

COMPOSITE VARIABLE – B



APPENDIX - B

VARIABLE DEFINITIONS

1) Population Density. Persons per square kilometre, where the number of persons is given by the total average population for 1977.

2) Birth Rate. The number of births per thousand population. The figures are attributed to the territorial unit of residence of the mother.

3) Mortality Rate. The number of deaths per thousand population. The figures are attributed to the territorial unit of residence of the deceased.

4) Unemployment Rate. The number of unemployed persons expressed as a percentage of the population.*

5) Age Index Figures are calculated from the index B/A, where A is the percentage of the population between 0 and 14, and A is the percentage of the population over the age of 64.

6) Employment in Agriculture Expressed as a percentage of the population.*

7) Employment in Industry Expressed as a percentage of the population.*

8) Gross Domestic Product Per capita. Values are expressed in Standard of Purchasing Power (SPA) units.

9) Net Migration Per thousand population.

* For the three employment variables, "population" refers to private households only, which represents about 97% of the total population.

INTERPOLATED VALUES

1) Birth Rate, Death Rate and Net Migration. The value for the Level I region of Germany, Neidershasen, was given to the eight Level II regions nested within it.

2) Unemployment Rate, Employment in Agriculture and Employment in Industry. Values for the three Level II regions of Denmark were interpolated from the value for Denmark. Values for the French Level II regions, Corsica and Provence, were interpolated from the supplied joint value.

3) Gross Domestic Product. All Level II regions of Germany received the value of their respective Level I region. (Five regions of Germany serve as both Level I and Level II in EUROSTAT´s regionalisation.)

For information about region names, see the "EUROSTAT 1977 Statistics" publication.