Spatial Analysis of Biodemographic Data

Analyse spatiale de données biodémographiques

Edited by/*Coordonné par*

Jean-Pierre BOCQUET-APPEL,
Daniel COURGEAU et Denise PUMAIN

1996

John Libbey EUROTEXT

INED

CONTENTS/*SOMMAIRE*

LIST OF AUTHORS/*LISTE DES AUTEURS*

Brigitte BACCAINI
Équipe Biographie et Peuplement
INED
27, rue du Commandeur
75675 PARIS Cedex 14
France
email: baccaini@ined.fr

Jean-Pierre BOCQUET-APPEL
CNRS, UMR 152
Musée de l'Homme
17, place du Trocadéro
75116 PARIS
France
email: bocquet@mnhn.fr

Jean-Luc BONNEFOY
Département de Géographie
Université de Pau et Pays de l'Adour
Av. du Doyen Poplawski
64000 PAU
France

Athanase BOPDA
OCISCA-ORSTOM
BP 1857
YAOUNDÉ
Cameroun

Chris BRANSDON
Depar[t] of Town and Country Planning
University of Newcastle
NE1 7RU NEWCASTLE
United Kingdom

Stephane BURA
LAFORIA
Université de Paris VI
75252 PARIS Cedex 05
France
email: bura@laforia.ibp.fr

Martin CHARLTON
Department of Geography
Daysh Building
University of Newcastle
NE1 7RU NEWCASTLE
United Kingdom

Daniel COURGEAU
Équipe Biographie et Peuplement
INED
27, rue du Commandeur
75675 PARIS Cedex 14
France
email: courgeau@ined.fr

Pierre DARLU
CNRS, Inserm U155
Épidémiologie génétique – Case 7041
Université Denis-Diderot
75251 PARIS Cedex 05
France

Jean-Michel DECROLY
Département de Géographie
Université libre de Bruxelles
Campus de la Plaine
Boulevard du Triomphe
B-1050 BRUXELLES
Belgique

Anna DEGIOANNI
CNRS (même adresse que P. DARLU)
et
Istituto de Genetica Biochimica
ed Evoluzionistica
Via Abbiategrasso, 207
I-27100 PAVIA
Italia

Stewart FOTHERINGHAM
Department of Geography
Daysh Building
University of Newcastle
NE1 7RU NEWCASTLE
United Kingdom
email: stewart.fotheringham@newcastle.ac.uk

Pierre FRANKHAUSER
IRADES
Université de Franche-Comté
30, rue Mégevand
25030 BESANCON
France

Claude GRASLAND
ÉQUIPE P.A.R.I.S.
13, rue du Four
75006 PARIS
France
email: grasland@paris.msh-paris.fr

France GUÉRIN-PACE
Équipe Concepts et méthodes en démographie : espace, temps société
27, rue du Commandeur
75675 PARIS Cedex 14
France
email: guerin@ined.fr

Tim HOLT
Department of Social Statistics
University of Southampton
Highfield
SOUTHAMPTON SO17 1BJ
United Kingdom
email: mdt@alcd.soton.ac.uk

Lucienne JAKOBI
CNRS, UMR 152
Musée de l'Homme
17, place du Trocadéro
75116 PARIS
France
email: jakobi@mnhn.fr

Norbert KOCH
Institut für theoretische Physik
Universität Stuttgart
Pfaffenwaldring 57/III
70550 STUTTGART
Deutschland

Hélène MATHIAN
ÉQUIPE P.A.R.I.S., CNRS
13, rue du Four
75006 PARIS
France
email: mathian@paris.msh-paris.fr

Alina POTRYKOWSKA
Institute of Geography
and Spatial Organisation
Polish Academy of Sciences
Krakowskie Przedmiescie, 30
00-927 Warsaw
Poland

Michel POULAIN
GÉDAR, Université de Louvain
Rue du Fayt, 28
5150 SOYE (Floreffe)
Belgique
email:

Denise PUMAIN
ÉQUIPE P.A.R.I.S., CNRS
et Université Paris I
13, rue du Four
75006 PARIS
France
email: pumain@paris.msh-paris.fr

Françoise ROLLAN
CESURBE
Maison des Sciences de l'Homme
Esplanade des Antilles
Domaine universitaire
33405 TALENCE Cedex
France

Céline ROZENBLAT
GIP RECLUS, Équipe P.A.R.I.S.
Maison de la Géographie
17, rue de l'Abbé-de-l'Épée
34000 MONTPELLIER
France

Lena SANDERS
ÉQUIPE P.A.R.I.S., CNRS
13, rue du Four
75006 PARIS
France
email: sanders@paris.msh-paris.fr

Timm SIGG
Institut für theoretische Physik
Universität Stuttgart
Pfaffenwaldring 57/III
70550 STUTTGART
Deutschland
email: sigg@theo2.physik.ini-stuttgart.de

David STEEL
Department of Applied Statistics
University of Wollongong
Northfields Avenue
WOLLONGONG NSW 2522
Australia
email: David_Steel@uow.edu.au

Mark TRANMER
Department of Social Statistics
University of Southampton
Highfield
SOUTHAMPTON SO17 1BJ
United Kingdom
email: mdt@alcd.soton.ac.uk

Christine VOIRON-CANICIO
Laboratoire d'Analyse spatiale,
URA D1476
Faculté des Lettres et Sciences humaines
98, boulevard Edouard-Herriot – BP 209
06204 NICE Cedex 3
France

Wolfgang WEIDLICH
Institut für theoretische Physik
Universität Stuttgart
Pfaffenwaldring 57/III
70550 STUTTGART
Deutschland

Gianna ZEI
CNRS (même adresse que P. DARLU)
et
Istituto de Genetica Biochimica
ed Evoluzionistica
Via Abbiategrasso, 207
I-27100 PAVIA
Italia

Preface/*Introduction*

The editors/*Les coordinateurs*

The characteristics of populations are considered less and less as being spatially homogenous. On the contrary, they are quite often regarded as depending on the real geographic space in which they are located. This is particularly the case for numerous demographic variables which are simultaneously biological ones, such as the age pyramid and degenerative diseases or congenital anomalies distributions (chromosomal defects, inborn error of metabolism), migrations and gene flows. This led researchers in social sciences to reconsider past approaches for the study of these biodemographic data and to set forth methods of analysis for their spatial distributions.

The intention of the organizers of the seminar 'Spatial Analysis of Biodemographic Data', which was held at the Sorbonne in Paris, in July 1995, was to give wide coverage to the recent advances in this domain. A meeting of about thirty researchers from different disciplines (demography, biological anthropology, geography, etc.) took place, during which new methods were discussed and compared with older ones. This book gives an overview of the three main sections which were developed during the seminar.

The first section deals with the analysis of demographic data, given their aggregation levels. It concerns multi-level analysis. Numerous models have been proposed in the past, using aggregated data observed in geographical zones or for groups of individuals, i. e. at a macroscopic level. More recently, analytical methods for biographies have permitted the use of 'desaggregated' data, observed at the level of the individual (microscopic), to explain diverse types of behavior. The search for the links between these two observation levels raises a number of methodological and conceptual problems. The modelization of spatial phenomena is renewed and brings forth important epistemological issues such as, the meaning of the diverse levels which are possibly observable.

The second section concerns the analysis of the spatial patterns and configurations which appear in biodemographic data (demographic, patronymic, genetic, epidemiologic), when they are mapped. During the processes of exchange, migration or diffusion, barriers of all kinds (physical as well as cultural) deviate the flow of information, give it a direction, slow it down or even stop it. This brings about loose meshes and major discontinuities in the spatial continuum of the variables. These spatial discontinuities generally delineate zones of greater homogeneity. When one analyzes spatialized data, before inferring putative causal variables, the first approach is to detect and then localize these zones of discontinuity. To this end, different techniques are advocated, which combine descriptive spatial statistics and image recognition.

The genesis of the observable structures in the geographic space is the subject of the third section in which the interactions brought about at a certain level of system organization are analyzed to understand the dynamics of these systems, but at a higher level. The self-

Les caractéristiques des populations sont de moins en moins souvent considérées comme spatialement homogènes. On estime au contraire qu'elles dépendent de l'espace géographique concret dans lequel elles se situent. Cela vaut en particulier pour de nombreuses variables démographiques qui sont simultanément des variables biologiques, telle la distribution des pyramides des âges et celle des maladies dégénératives ou cliniques, les migrations et les flux génétiques. Cela conduit de nombreux chercheurs en sciences sociales à reconsidérer les anciennes approches de ces variables biodémographiques et à mettre en place des méthodes d'analyse de leurs répartitions spatiales.

Le séminaire sur l'« Analyse spatiale des données biodémographiques » qui s'est tenu à Paris en juillet 1995 tentait de faire le point sur les avancées récentes dans ce domaine. Il a permis la rencontre d'une trentaine de chercheurs de différentes disciplines, qui ont confronté et discuté leurs méthodes d'analyse, avec d'autres, plus anciennes. Cet ouvrage donne une vue d'ensemble des trois grands thèmes qui ont été développés lors de ce séminaire avec, en contrepoint, des études de cas plus détaillées.

Un premier thème porte sur l'analyse des données démographiques selon le niveau d'agrégation auquel on se situe. Il s'agit de mettre en place des analyses multi-niveaux. De nombreux modèles ont été proposés dans le passé, qui utilisent des données agrégées, observées dans des zones géographiques ou pour des groupes d'individus, c'est-à-dire au niveau macroscopique. Plus récemment, les méthodes d'analyse des biographies ont permis d'utiliser des données « désagrégées », observées au niveau individuel (ou microscopique), pour expliquer divers types de comportements. La recherche des liens qui existent entre ces deux niveaux d'observation soulève actuellement de nombreux problèmes d'ordre méthodologique et conceptuel. La modélisation des phénomènes spatiaux en sort renouvelée et amène à poser des questions épistémologiques importantes, notamment quant à la signification des divers niveaux qu'il est possible d'observer.

Un second thème porte sur l'analyse des traces et des configurations spatiales qui apparaissent du fait que les données biodémographiques (démographiques, patronymiques, génétiques, etc.) ne sont pas homogènes à travers l'espace. Au cours de processus comme les échanges, les migrations, ou les diffusions, par exemple, des barrières de toute nature (physiques ou culturelles) orientent la direction des flux, les freinent, voire les arrêtent, créant des mailles lâches, des discontinuités dans le continuum spatial des variables. Ces discontinuités séparent généralement, pour les variables considérées, des zones de plus grande homogénéité. Lorsque l'on analyse des données spatialisées, la première approche consiste à repérer ces zones de discontinuité, avant de tenter d'en inférer les variables causales sous-jacentes. Différentes techniques sont élaborées pour permettre de détecter ces zones, elles combinent la statistique spatiale et l'analyse d'image.

Un troisième thème s'intéresse à la genèse des structures observables dans l'espace géographique. Il considère les interactions qui se produisent à un certain niveau d'organisation des systèmes pour comprendre la dynamique

organization theories offer a new frame to interpret and model the raising of territorial entities, like the cities (from the interaction of the urban actors), the regions, or the urban systems (from the interactions of the towns). The analysis of the transition process, from one organization level to the other, the emerging qualitative structures which makes relevant the change of the observation scale, are among the important issues raised by the evolution of the populations through time and space.

This volume offers an overview of the different approaches used to improve our understanding of spatial distribution of biodemographic data, but also provokes general relexion on the complexity of spatial interactions produced through time and with information units as different as the individual, the family, the neighborhood and the city, the region, the country and the continent. These approaches pave the way to new avenues in human sciences and we may surmise that the contributions selected here augurs well of the future in this field.

Acknowlegments

We would like to thank The International Union for the Scientific Study of the Population (IUSSP) and Jacques VALLIN, its representative at the Seminar, the *Institut National d'Études Démographiques (INED)*, the French ministry *de l'Éducation Nationale, de l'Enseignement Supérieur, de la Recherche et de l'Insertion Professionnelle* and the *Centre National de la Recherche Scientifique (CNRS)*, whose help permitted us to organize this seminar and to publish this book. Many thanks to the *Université de Paris IV* and his president Jean-Pierre POUSSOU, who invited this seminar to take place in the Sorbonne. An anonymous referee must also be thanked for his valuable remarks on a first version of the manuscript. Finally we wish to thank Cyril COURGEAU who redraw the graphs, figures and maps and Hella COURGEAU who produced and layed out this volume.

de ces systèmes à un niveau supérieur d'organisation. Les théories de l'auto-organisation proposent ainsi un nouveau cadre pour interpréter et modéliser l'émergence d'entités territoriales comme les villes (à partir des interactions entre les acteurs urbains), les régions, ou les systèmes de villes (à partir des interactions entre les villes). L'analyse des processus par lesquels s'effectue le passage d'un niveau à un autre, l'émergence qualitative des structures qui rend pertinent le changement de l'échelle d'observation, sont parmi les grands problèmes posés actuellement à propos de l'évolution spatio-temporelle des populations.

Cet ouvrage permet non seulement de dresser une vue d'ensemble de la variété des méthodes qui permettent de mieux comprendre la répartition spatiale des données biodémographiques, mais ouvre une réflexion plus générale sur la complexité des interactions spatiales qui se produisent également au cours du temps, avec des unités aussi différentes que l'individu, la famille, le quartier, la ville, la région, le pays, etc. Ces nouvelles approches ouvrent un domaine encore peu exploré en sciences humaines, mais dont on peut penser que les développements présentés ici laissent bien augurer de sa richesse à venir.

Remerciements

Nous remercions l'Union Internationale pour l'Étude Scientifique de la Population (UIESP) et Jacques VALLIN, son représentant au séminaire, l'Institut National d'Études Démographiques (INED), le ministère de l'Éducation Nationale, de l'Enseignement Supérieur, de la Recherche et de l'Insertion Professionnelle et le Centre National de la Recherche Scientifique (CNRS), qui ont permis l'organisation de ce séminaire et la publication de cet ouvrage. Nous remercions également l'université de Paris IV et son président Jean-Pierre POUSSOU d'avoir accueilli ce séminaire à la Sorbonne. Nous tenons également à remercier un rapporteur anonyme pour ses commentaires et critiques sur une première version de cet ouvrage. Notre gratitude va enfin à Cyril COURGEAU qui a remis en forme les graphiques, figures et cartes de l'ouvrage et Hella COURGEAU qui en a assuré l'élaboration et la mise en pages.

Towards a Multilevel Analysis in Social Sciences/ *Vers une analyse multi-niveaux en sciences sociales*

Daniel COURGEAU

The research on relationships between different biodemographic characteristics, seems *a priori* easier to undertake when using individual data, obtained from surveys or specific population registers. In this case, the analysis is usually undertaken by using logit, probit, log linear or regression models with cross-sectional data, or by using event history analysis with longitudinal data. However, the relative scarceness of such surveys and population registers or the wish to put into evidence relationships at a different level of aggregation, may lead the researcher to work on aggregate data sets, considered at different levels: parishes, communes, regions, etc.

This occurs particularly when, in order to get the essential data for the analysis, it is necessary to work on different data sets: vital registration, censuses, and specific surveys. In this case it is no longer possible to reconstruct individual data, but it is always possible to analyse the relationships between rates or indexes which are estimated on each of the considered areas, using different regression models.

The use of one or another kind of data is grounded on different hypotheses, which we now have to take into consideration before approaching a multilevel analysis.

Distinction between individual and aggregate level data

When the analysis is undertaken at the individual level, its aim is to give a model for a demographic or biologic behaviour related to different characteristics of individuals which are supposed to affect such behaviour. Underlying this aim is the hypothesis that only individual characteristics will affect personal behaviour. It is easy to see how such an hypothesis can be found guilty of the *atomic fallacy*, as we miss the context in which individual behaviour occurs. In fact, such a context may play a role on individual behaviour and it seems misleading to consider an individual who is acting according to his own characteristics without consideration of the restraints imposed by the society and the environment in which he is living.

Conversely, when the analysis is undertaken at the aggregate level, such an approach at least permits one to take into account the effect the characteristics of the studied groups may have on group behaviour. The aggregate characteristics that are measured, can approximately represent different restraints that the diverse groups may impose on their members. On the other hand, such an approach lays one open to the charge of what is usually called *ecological fallacy* which fails to recognize that individuals do act, not aggregates. In this way an ecological correlation in which the statistical object is a group of persons, is almost certainly not equal to an individual correlation in which the statistical object is indivisible. This leads to an aggregation bias of

individual level relationships. However, such a bias may be more especially reduced as the within-area variance of each characteristic is lower than the between-area variance, i.e. the ecological one. Usually, such an hypothesis is far from being verified, so that it is always difficult, to draw individual conclusions from aggregate data when individual data are not available (which is unfortunately often the case).

In view of this dilemma, some researchers recently explored the possibility of working simultaneously on different levels of aggregation (Jones, 1993; von Korff *et al.*, 1992; Courgeau, 1995). Such an approach will cause us to re-examine the problems we met when working at only one aggregation level and to display the hypotheses necessary to undertake a *multilevel analysis*. We will show, hereafter, some new tools which are being used to carry out such an analysis.

Problems and hypotheses for multilevel analysis

As soon as one tries to undertake an analysis introducing different levels of aggregation, a number of distinctions between response variables and predictor variables are to be introduced. A first kind of variable corresponds to *individual characteristics* – biological as well as demographical – blood pressure, height, birth of a child, death of an individual, etc. In this case, it is easy to see how to aggregate such characteristics to get a new variable at any aggregation level: mean blood pressure, mean height, number of births, or birth rate, number of deaths or death rate, etc., measured on any arbitrary area. However, there are other characteristics that have no corresponding attribute at the individual level as *contextual properties*: population density, for example. These characteristics however can be aggregated at different macro-levels: population density of communes, parishes, regions, etc. Finally, there are some other *collective characteristics* which are defined at a given level of aggregation. For example, the hardness of the water is identically defined in its whole drainage basin and affects the whole population living in this district. It constitutes a measure of the individual exposure of all the members of the district. If we try to change the geographical areas, it may always be possible to define a mean hardness of the water, on weighting by the different subpopulations submitted in each different basin. Yet such a measurement seems too arbitrary and of little interest to be considered, so that it seems necessary to consider only the level at which it is defined.

We have also to draw a distinction between *biological characteristics* of individuals, for which an individual measurement will be the best one, and other *socio-demographic characteristics*, for which there is no reason to choose between an individual measurement or a more aggregate one. As a matter of fact, to analyse socio-demographic characteristics, event

history analysis can explain a time dependant individual characteristic (migration rate, fertility rate for a given rank, mortality rate, etc.) by different other characteristics of the same individual (to be a farmer, a foreigner, etc.) which can also be time dependant. Similarly, econometric regression models can use aggregate characteristics. For example, regional out migration rates can be explained by the percentages of individuals in different situations (percentage working in agriculture, percentage of foreigners, etc.)

Consequently, it appears that an important problem in multilevel analysis will be to find what the relationships between these different models are and what new contribution can be taken from such a synthetic approach compared with event history analysis and regression analysis undertaken separately.

Under these conditions, it may be agreed to unify methods of analysing biological data and socio-demographic characteristics. In order to explain an individual characteristic, it seems better to simultaneously introduce independent variables measured at the individual level and characteristics measured at different levels of aggregation. It may be the same characteristic measured at individual and diverse aggregation levels (individual working in agriculture / proportion of farmers in each region). They may also be characteristics defined only at aggregate levels (population density in each parish / population density in each region). Finally, they may be characteristics which only exist at a given aggregation level (hardness of the water distributed in a given district).

Conversely, one may explain an aggregated characteristic by different variables measured at the same aggregation level or at different regrouping levels. If in order to explain aggregated data, it is not possible to come back to individual explanatory variables, other summaries than the regional proportions may be used to improve the fit of the model. In any case, such an approach introduces an important asymmetry between individual and more aggregate levels.

Now that we have better defined the field to explore, it is important to establish if the methods of measurement generally used in social sciences can give us the all of the useful information to undertake such an analysis.

Amongst the exhaustive measures, population register data are a serviceable source of multilevel data. When these population registers are able to cover a great number of fields in the social life, like the Danish registers, they are in a favourable position. Unfortunately, the number of countries having such population registers is very minimal and the field covered is often restricted in many of these countries: births, marriages, internal migrations and deaths are the most frequently registered events. This will permit only limited studies with such a little number of registered events.

Under these conditions, it seems useful to undertake more detailed surveys, even in countries with population registers. Such surveys based on samples which are often of small size, will have to collect not only individual information, but also information on different characteristics of areas in which these people live: hardness of the water, proximity to a market or health center, etc. This will permit to set up the whole set of individual and aggregate characteristics in order to be able to undertake a multilevel analysis.

The usual methods of analysis can be applied to such data, with some changes: regression, logit, probit models and others, when the characteristics are not time dependant; event history analysis models when they are time dependant. However, more specific new multilevel statistical models need to be developed to analyse such data (Goldstein, 1995). It also seems necessary, as these aggregation problems occur in each social sciences to compare their different approaches in order to be able to determine human behaviour correctly.

Also, this multilevel approach raises new epistemological questions, especially about the status to be given to aggregate characteristics. Are they a reflection of the social organization in which we are living, whereas individual freedom would better appear under individual characteristics? If such is the case, what meaning can we give to a great number of intervening aggregation levels? Would it not be better to display privileged levels which would be integrated in a more general theory? Lastly, how can we link different levels such as the individual, the family, and the neighborhood, the city, the region, and the whole country?

Even if these problems are far from being solved, a reflection on this theme would be very important to undertake in the future.

Some new tools

A number of research workers in different social sciences (biology, demography, human geography, social statistics, etc.) met to take stock of recent progress in multilevel analysis.

Stewart Fotheringham *et al.*, apply themselves to analyse a synthetic characteristic, for example population density, measured at different aggregation levels. In this case, we could not have an individual measure, as density is only defined at an aggregate level. In return, some methods, that permit to explore spatial non-stationarity are explored here and compared to classical use of regression with spatial data investigating the relationships between population density and a series of attributes of the physical landscape.

On the contrary, Tim Holt *et al.* apply themselves to estimating individual level regression coefficients, when only aggregate data for geographic areas are available and, with a small amount of areas, an individual level information on a set of "grouping variables" which come from another data set. Their theoretical investigation provided new results on the problem of ecological fallacy. Also, their empirical investigation on biological data using blood pressure show the contribution such an approach brings to the comparison between individual and aggregate data.

Michel Poulain then tackles the problem of collecting multilevel data, in order to undertake socio-demographic studies. He shows the necessity to modify data input to permit such analyses. The use of population registers, which are centralized and computerized, is described here and their potentialities are shown with detailed example cases.

Brigitte Baccaïni and Daniel Courgeau approach the question of using the same socio-demographic data at an individual level and at aggregate level. They show first from a theoretical approach, the links between the estimated parameters of regression, logit and event history models which may be verified if no aggregation effect exists. They used Norwegian population register data. They gave an estimation of the different model parameters and concluded that the effect of individual characteristics is almost independent of the effect of aggregates characteristics. The two levels each brought an independent information on the characteristics studied here, the out-migration of Norwegian regions.

Towards a theory of multilevel models

The different advances explored in the papers from this session show the potential wealth of multilevel models and, simultaneously, the need of a coherent theory for these models. First, such a theory has to go beyond the concepts of ecological fallacy and of atomistic fallacy. In fact, the ecological fallacy is the result of an improper inference on individual-level associations based on associations measured only at the aggregate level. Similarly, the atomistic fallacy is the result of an incomplete specification of the different characteristics acting on individual behaviour. As soon as one acknowledges the importance of both individual and environmental variables in determing individual level behaviour, the ecological and atomistic fallacy no longer exists.

Furthermore, such a theory needs to go beyond the study of individual behaviour, which tries to explain it by using different characteristics measured at different aggregation levels: it may study the behaviour suited for each aggregation level which will simultaneously further explain each of them in relationship to the others. So that individuals acting in a given community can become aware of a problem

which affects the collectivity and lead thus, to political action taken at a more aggregated level. Conversely, these political measures will affect individual behaviour, and so on.

Finally an important problem occurs, from the possible existence of a great number if not an infinity, of existing aggregate levels. It seems evident that some of these levels may be more important than others, but the precise criteria and the methods to detect them are, for the moment, waiting to be settled. Evidently, it is easy to consider some aggregation levels with an administrative definition (city, commune, region, etc.), but it is not obvious to say if they are the most relevant ones to be considered and if other levels, more difficult to define, such as a community, are not of greater interest. These aggregation levels may also change according to time and some new aggregation levels may appear which will have to be explained.

In conclusion, it may be said that such a theory is entirely pending and that its epistemological basis, its methods of measurements, and of analysis are not yet stated. In this paper we provoke thought about a theory of multilevel models, which will try to synthestize human and biological sciences and we hope that forthcoming research will bear out its importance.