LIVES
Swiss National Centre of Competence in Research

FNSNF
Swiss National Science Foundation

Courgeau, D. (2016)

Do different approaches in social science lead to divergent or convergent models?

in G. Ritschard & M. Studer (eds), Proceedings of the International Conference on Sequence Analysis and Related Methods, Lausanne, June 8-10, 2016, pp 7-30.

UNIVERSITÉ DE GENÈVE

FONDATION POUR L'UNIVERSITÉ DE LAUSANNE

UNIVERSITÉ DE GENÈVE
GENEVA SCHOOL OF SOCIAL SCIENCES

UNIL | Université de Lausanne
Institut des sciences sociales

Do different approaches in social science lead to divergent or convergent models?

Daniel Courgeau

Institut National d'Etudes Démographiques (INED)

Abstract Sequence analysis is essentially an exploratory tool and needs to be complemented with other modelling approaches when it comes to testing hypotheses or studying the dynamics that drives the trajectories. This paper will first explore some of these tools: event-duration models which lead to event history analysis; event-sequences models which lead to sequence analysis; multiple level models which lead to multilevel analysis; social network models which lead to multilevel social-network analysis; models based on individual agents which lead to agent-based analysis. It then shows that these models can be classified under some more general concepts: the *statistical individual* concept covers event history and sequence analysis; the *statistical network* concept covers multilevel and social-network analysis. Only the agent-based analysis seems to escape from these concepts as it models theoretical ideas rather than data. However as it remains at the individual level it is too reductionist to explain social behavior. It seems then necessary to set up a more robust research program for demography. This research program may follow the induction's way given by Bacon in searching for the structure of the studied phenomena and the interactions between the networks created by people. Such a program will be able to lead to a convergence of these different models.

1 Introduction

From it's inception by Graunt in 1662, the scientific study of population called by Petty (1690) *political arithmetick,* paved the way for around 200 years for demography, epidemiology, political economics, and more generally for population sciences. During this period a *cross-sectional* approach was followed for which social facts of a period exists independently of the individuals who experience them, and can be explained by various characteristics of their society. After the end of World War II, population scientists took a new view on these facts, which introduced the individual's lived time. This *cohort analysis* approach considered that the occurrence of a given event, during the life of a generation or a cohort, can be studied in a population which preserves all its characteristics and the same charac-

teristics for as long as the phenomenon manifests itself (Courgeau, 2007). This approach was however submitted to very restrictive conditions (Courgeau and Lelièvre, 1994), and leads to the more recent approaches which we will present more thoroughly in this paper.

Social scientists today use various methodological approaches that perform often complementary but sometimes divergent tasks. We shall first briefly describe the main methods used in population science, emphasizing their potential convergences as well as their divergences.

From this comparison, we shall try to identify the conditions that would allow a synthesis of the approaches through an analysis of a more epistemological nature regarding an inductive construction in the Baconian sense (1620). This method of induction[1] consists of discovering the principles of natural or social processes by way of experimentation and observation. It rests on the requirement that without these principles the properties observed would be different (Franck, 2002).

2 Research Areas

We shall present and mainly discuss here the five main approaches by event duration, event sequence, multiple levels, network and agent-based decisions, used in population science.

2.1 An approach by event duration

This first approach made its debut in social sciences in the early 1980s, more than thirty years after the introduction of longitudinal analysis. However, it was already in circulation earlier, particularly among statisticians. We can trace its origin to the notion of martingale, used by Ville in 1939 and Doob in 1953. In 1972, Cox proposed the joint use of life tables and a regression model. In 1975, Aalen suggested the use of counting process theory for the joint analysis of several events that an individual could experience simultaneously. In 1980, the analysis by Aalen et al. of the interaction between events in an event history introduced the approach into the field of population sciences.

This approach rests on robust mathematical and statistical foundations, which permit to establish risk factors and to treat censored observations. They are presented in statistical books by Kalbfleisch and Prentice (1980), Cox and Oakes (1984), Andersen et al. (1991), and Aalen et al. (2008). They make it possible to analyze changes of state, however diverse, and to demonstrate the role of many individual characteristics that can change over time during such transitions. The

[1] Induction is not taken in the sense of Mill (1843) and his followers, i.e. generalization from particular facts. In Bacon's sense, induction designates the complete research process.

application of the method in demography (see for example Courgeau and Lelièvre 1992) brought fresh progress in that field. Many other social sciences adopted it as well, including epidemiology, biostatistics, sociology, econometrics, actuarial sciences, and medicine.

The event-history approach eliminates the need for the overly restrictive hypotheses of longitudinal analysis while maintaining the individual point of view. Individuals can be tracked over a part of their entire lifetime, typically by means of a retrospective or prospective survey. It focuses on the duration between different events occurring in a person's life, and its application requires special surveys. For example, in 1981 the "triple event history" survey (currently called 3B, see Courgeau, 1999) allowed the simultaneous analysis of family events, occupational events, and migration events occurring over a lifetime up to the survey date for cohorts born between 1911 and 1936. As censored observations can be treated without problem by this approach, the persons who were always in the labor force at the time of the survey (four fifths) can be studied for their occupational history in the same manner than those who were retired (one fifth).

It basically relies on semi-parametric methods, which, while preserving a non-parametric vision of the duration between events, use parameters to describe the effects of personal characteristics (Courgeau and Lelièvre, 1992).

However, the event-history approach did pose a certain number of problems, to which we now turn.

The first problem is that of unobserved heterogeneity. How does unobserved heterogeneity affect the estimation of parameters of observed characteristics? To help us answer the question, we have an important result obtained by Bretagnole and Huber-Carol in 1988 but overlooked by some users of these models. The two authors showed that, in a Cox model, when the omitted characteristics are independent of the observed characteristics, the omission has no impact on the sign of the estimated parameters, reducing only their absolute value. Thus, if the effect of a characteristic is found to be fully significant, the introduction of unobserved characteristics will merely strengthen that effect. Conversely, a characteristic that does not have a significant effect may have one when the omitted characteristics are introduced. We need to be aware of this risk.

When observed and omitted characteristics are connected, the situation is more complex. It may be tempting to introduce this heterogeneity as a particular type of distribution, which Vaupel et al. called *frailty* in 1979. When we have information on the distribution, its introduction is entirely legitimate. The problem is that we typically do not know this distribution, and that it is often chosen for no other valid reason than convenience. In such circumstances, some estimates may even change the sign of certain parameters, as Trussell and Richards showed in 1987, while a model without frailty avoids this problem.

We therefore totally agree with Aalen et al. (2008), who, in their extensive studies on stochastic processes, have tried to identify individual frailty:

> As long as there is no specific information about the underlying process and observations are only made once for each individual, there is little hope of identifying what kind of process is actually driving the development.

4

Indeed, for the analysis of non-repetitive events, there is only one model without observed heterogeneity, but an infinity of models with unobserved heterogeneity. Their estimates differ, but they display an identical fit with observed data (Trussell, 1992). By contrast, if we are analyzing repetitive events—such as successive births or migrations—we have the option of estimating multilevel models that allow the introduction of unobserved heterogeneity, which reflects the multiple events experienced by every individual. We shall present these multilevel models later.

The second problem concerns the concept of probability used. Apart from Kalbfleisch and Prentice, most of the earlier-mentioned statisticians who developed the method chose an objective probabilistic approach, which places certain constraints on the expected results of an analysis. Could an epistemic approach enable us to lift many of these constraints? We cannot give a full description of the probabilistic approach here, such as in Courgeau (2012), but we can elaborate on the constraints linked to statistical inference.

The purpose of statistical inference is to optimize the use of the incomplete information available in order to take the best decision. Statistical inference will therefore consist in providing an analysis of a past phenomenon and a prediction of a similar phenomenon to come. The first point is important for sciences such as demography or epidemiology, which must analyze human behavior. The second point is crucial for sciences, such as medicine, or those focusing on public health which aim to produce the best possible forecast of the outcome of a treatment course or a decision on the best policy to implement. Statistical inference notably leads to testing various hypotheses about the phenomena studied.

Objectivist methods, also called frequentist methods, seek to verify whether a given factor does or does not affect the phenomenon studied, and this brings us to the notion of statistical test. This means treating the sample under analysis as one possible selection from an infinity of other samples that we extract from a population also assumed to be infinite. When we assign a confidence interval of, say, 95% to a parameter estimated on this sample, we might conclude that the probability of the unknown parameter lying in the interval is 0.95. In fact, however, the objectivists tell us that this conclusion is wrong. All we can state is that if we draw an infinity of new samples, then the new estimated parameters will lay in that interval 95% of the time. As Jeffreys wrote in 1939, when examining various definitions of objective probability:

> The most serious drawback of these definitions, however, is the deliberate omission to give any meaning to the probability of a hypothesis. All they can do is to set up a hypothesis and give arbitrary rules for rejecting it in certain circumstances.

That is exactly what happens with statistical tests. Similarly, the use of frequentist methods for prediction will consist in taking the parameters estimated, for example, by means of maximum likelihood and introducing them into the distribution function of the new observation. But this will not allow us to factor in the uncertainty of the parameter estimation, and will lead to an under-estimation of the variance of the predicted distribution.

That is why Jeffreys himself showed that if we accept that a probability is never a frequency—in other words, if we adopt the epistemic framework—then a 95% confidence interval truly means an interval in which the statistician rightly believes that the unknown parameter may lie with a probability of 0.95. Moreover, this approach enables us to solve the prediction problem, for which the objective approach could provide only an approximate solution. All we need to do is calculate the "*posterior predictive distribution*" of a future observation from the initially observed data, which are known. What we obtain is not a value, as with the objectivist method, but a distribution whose variance will now be calculated correctly.

We have not described in detail all the advantages of using an epistemic method, but they have led a number of authors to propose it for event-history analysis, especially when the sample studied is small: see for example the book published by Ibrahim et al. in 2001. However we will see in *2.3* that another way of making statistical inference is possible with epistemic probability.

The last problem we would like to address is that of the risk of *atomistic fallacy* involved in this approach. If we can draw on all individual characteristics to explain a behavior, we shall overlook the context in which the behavior occurs. In fact, when using a cross sectional approach the researcher introduced only the characteristics of the society to explain social facts. This aggregate approach was on the contrary under the risk of *ecological fallacy*, as Robinson (1950) so clearly demonstrated: he showed that the correlations between two characteristics measured in binary mode on individuals, or by proportions applied to different geographic segmentations, generally diverged. We will se later how to solve this difficulty.

2.2 An approach by event sequences

We can trace the origin of sequence analysis in computer science as used by Levenshtein (1966); then in molecular biology for the study of DNA and RNA sequences as used by Levitt (1969). It was introduced later in the social sciences, with the work of the sociologist Abbott (1983, 1984) in order to study social processes which occur by whole sequences generally during a long period of time.

However, this approach in social science rests on less robust mathematical and statistical foundations than event history analysis. Its main object is to describe whole sequences (ordered list of elements) in terms of types that reflect socially meaningful trajectories experienced by subjects (individuals or more general entities like stimuli in psychology or artifacts in archaeology). It follows a two-step approach. First it tries to compute a distance between sequences under some operations (insertions / deletions called "indels" or substitutions) with a given cost for each operation. The main used metric is called Optimal Matching (OM), but we will see later that many other methods to compute these distances may be used. Then in a second step, using cluster analysis, it is possible to detect types of sequences, regrouping the whole set of subjects into exclusive and mutually exclu-

sive categories. Cornwell (2015) gives a more detailed description of these methods. A great number of social sciences adopted it: sociology, demography, psychology, economics, anthropology, political science, linguistics, etc.

The sequence approach permits to turn from the cross-sectional research of causes in Durkheim's sense (1895), to an emphasis on contexts, connections and events which Abbott (1995) called a quiet revolution in social science. The surveys used to track subjects over their life time are very similar to event history surveys, with an emphasis on the observation of whole processes without censoring. Their goals, however, are very different: while event history analysis seeks the causes of the studied phenomena, sequential analysis explores the paths followed without offering reasons for the underlying processes that generate them (Robette and Bry, 2012). So that individual characteristics need not to be recorded for this analysis, out of their event sequences and their characteristics before the analyzed sequences. For example the 2001 "Event histories and contact circle" survey made by Lelièvre, on a sample of cohorts born between 1930 and 1950, following the example of the "triple event history" survey but more detailed, permitted to apply sequence analysis to the professional trajectories of mothers and their daughters in order to compare them (Robette et al. 2012).

Contrary to event history relying mainly on semi-parametric methods, it basically relies on non-parametric methods which make no assumption about the process underlying the life course. Its aim is to explore and describe the course of events as a whole, without trying to focus on the risk of experiencing events and their determinants. There are also some recent Bayesian extensions of social sequence analysis (Bolano, 2014), through Hidden Markov models similar to biological approaches (Liu and Logvinenko, 2003).

However, this analysis poses some new problems, different from those posed by event-history analysis, to which we now turn.

The first problem lies in the metric used, particularly in the use of OM methods for social sciences. As we have said this approach was imported from information theory and molecular biology. For these disciplines the hypotheses lying in their foundations have been shown to be plausible and of wide applicability: the model proposed by Levitt (1969) for transfer ribonucleic acid confronted with experimental observations was in good agreement with them and theories about chemical processes give a strong support to these methods. However in social sciences the structure of sequences appears to be much more complex. As Wu (2000) said:

> Part of my skepticism stems, in part, from my inability to see how the operations defining distances between trajectories (replacements and indels) correspond, even roughly, to something recognizably social.

For example giving the same cost to a transition from unemployment to employment than from employment to unemployment seems highly implausible. This skepticism was more clearly demonstrated by Bison (2009), while the use of OM techniques has multiplied. He clearly shows, by means of simulations, that varying the substitution and indels costs may produce inconsistent results. This may lead to find regularities even when they do not exist (Bison, 2014).

In order to solve this problem a number of generalizations of OM method were proposed: variable substitution costs, different distance measures, spell-adjusted measures, non-alignment techniques, monothetic divisive algorithm (MDA), etc. See Cornwell (2015) for more detail on these improvements.

However, while the number of distances and costs measurements increases the problem of their comparability becomes more and more important. While comparisons exist between few different metrics, using empirical data, the only study comparing a large number of metrics, using a reasoned set of artificial sequences, was made by Robette and Bry (2012). They did not try to find the best metric but "rather to unravel the specific patterns to which each alternative is actually more sensitive". Even if they found some differences between the results of these metrics, "the main patterns they conceal will be uncovered by most of the metrics". However the differences exist and the inconsistent results found by Bison let the problem of the used metric largely unsolved.

The second problem lies in the use of cluster analysis for detecting classes of sequences. This method of classification was used long before sequence analysis, as it was already the title and the subject of a book written by a psychologist (Tryon, 1939) for manual calculation. When computers developed, they permitted not only an increase in the use of cluster methods but also a development of an increasing number of techniques to detect these groups. Simultaneously a great number of problems associated with clustering techniques appeared.

A paper by Everitt (1979) developed some of them, which we will present here shortly. One of the most important criteria for a good cluster solution lies in the choice of the number of groups that should exist in a given study. Unfortunately when the classification criterion is plotted against the number of groups, in the majority of cases, no "sharp step" permits to determine the best number of classes which remains entirely subjective. Other attempts to solve this problem let it unresolved. The assessment of the validity and stability of the clusters found by different techniques poses also problems. As there are many reasons leading different analyses to arrive at different sets of clusters, it is important to show the validity of such analyses and more importantly the validity of the hypotheses lying behind them. Unfortunately there are few validity tests of these different approaches, and fewer tests of their social meaning (Cornwell, 2015). We can cite Byrne and Uprichard recent comment (2012) on these problems: "Although written in the late 1970s, actually many of the 'unresolvable problems' raised in Everitt's article are still problems today".

The emphasis on context, connections and events leads sequence analysis to abandon regression methods and to consider the research of causes as obsolete. This leads to a third problem: "could clusters be an artifact of not controlling, say, for an observed variable?" (Wu, 2000). If the problem of unobserved heterogeneity was important for event history analysis, here even observed heterogeneity leads to difficulties. While sequence analysis attempts to approach trajectories as a whole, it is only possible to introduce characteristics measured before the starting of the analyzed trajectory. Introducing characteristics measured later or time dependent ones will lead to many conceptual issues and these characteristics are very rarely incorporated. However we will see in the part on synthesis that new at-

8

tempts to combine event-history and sequence analysis may permit to solve this difficulty (Studer et al., 2016; Rossignon et al., 2016).

A fourth problem is linked to the fact that sequence analysis cannot handle censored observations, contrary to event-history analysis: it views its subject of analysis as a single unit at its completion and can only analyze fully observed trajectories, leaving aside the partially observed ones. Such a limitation will let aside incomplete trajectories and will only permit a study of the past. For example, as the age at retirement was 65 years at the time of the 3B survey, if we want to make a sequence analysis of professional life history, we could only be able to make it on people born between 1911 and 1916 only, while the survey covers people born between 1911 and 1936. Like the previous problem similar authors are trying to circumvent this difficulty, accentuating the similarity between the analyses of event duration and event sequences.

Sequence analysis permits to describe the trajectories in terms of types or classes which are considered to reflect socially meaningful patterns experienced by subjects. However, the meaning of these patterns appears to be not so clear. First as one individual is allocated to one and only one type, this leads to a very narrow classification, while we know that an individual may in fact be allocated to a great number of groups such as family, business firms or organizations, contact circles, etc. These groups are real entities while the types given by a sequence analysis may be questioned. Second what are the grounds to believe in the existence of such types? Abbott and Tsay (2000) argue that sequence methods "would find this particular regularity because people in particular friendship networks would turn up in grouping of similar fertility careers". Their argument however presumes that data on friendship networks are available simultaneously with data on the fertility history of the same people. Unfortunately as far as I know, we have no examples showing the congruence of cluster results with friendship networks.

More recently a number of authors have similarly argued that network analysis may be a valuable tool to solve a number of these problems. For example Bison (2014) proposes to convert individual sequences into network graphs. Even if this method permits "to bring out career patterns that have never previously been observed", it has important limitations. As he said, the main one

> that creates methodological and philosophical problems is the annulment of individual sequences. … Everything is (con)fused to form a different structure in which the individual trajectories disappear to make space for a 'mean' trajectory that describes the transitions between two temporally contiguous points.

If we want to remain with the fundamental description of sequence analysis given previously, this point is really confusing. However Cornwell (2015) goes further and devotes a whole chapter on *Network methods for sequence analysis*. Even if some methods used in network analysis may be useful in sequence analysis, it is important to say how the object of each approach is different. For sequence analysis as we have already said its main object is to understand a life history as a whole and to identify regularities and structures. For network analysis, as we will see in *2.4*, the main object is to understand the relations between entities (individ-

uals, or more general levels of collective agency) and to see how changes at each level drives the evolution at other levels. We will try to find a solution to this problem in the final synthesis in part *3* of this paper.

We will have to see now how to introduce a more complex approach.

2.3 From a contextual to a multilevel approach

While the two preceding analyses operated at a given aggregation level, contextual and multilevel analyses introduce the effects of different levels on human behavior. It derived from the hierarchical models used in biometrics and population genetics since the late 1950s (Henderson et al., 1959). Their application and generalization to the social sciences came later—in sociology with Mason et al. (1983) and in education science with Goldstein (1986).

The simplest solution for introducing a contextual dimension is to incorporate into the same model the individual and aggregate characteristics of the groups involved, as Loriaux showed in 1989. We can now grasp the difference between this approach, which uses aggregate characteristics to explain an individual behavior, and the aggregate approach, which explained an aggregate behavior by equally aggregate characteristics.

We can thus eliminate the risk of *ecological fallacy*, for the aggregate characteristic will measure a different construct from its equivalent at the individual level. It no longer acts as a substitute, but as a characteristic of the sub-population that will influence the behavior of a member of that sub-population. Simultaneously, we remove the *atomistic fallacy*, as we take into consideration the context in which the individual lives. We may ask, however, if the inclusion of the aggregate characteristics provides an entirely sufficient representation of that context: as we shall see, it will be necessary to take further steps in a fully multilevel analysis.

In fact, the use of contextual models imposes highly restrictive conditions on the formulation of the log-odds (logarithm of relative risks) as a function of characteristics. In particular the models assume that the behaviors of individuals within a group are independent of one another. In practice, the risk incurred by a member of a given group more likely depends on the risks encountered by the group's other members. Overlooking this intra-group dependence generally biases the estimates of the variances of contextual effects, generating excessively narrow confidence intervals. Likewise, these log-odds, for individuals in different groups, cannot vary freely but have restrictive constraints imposed by the model used (Loriaux, 1989; Courgeau, 2004).

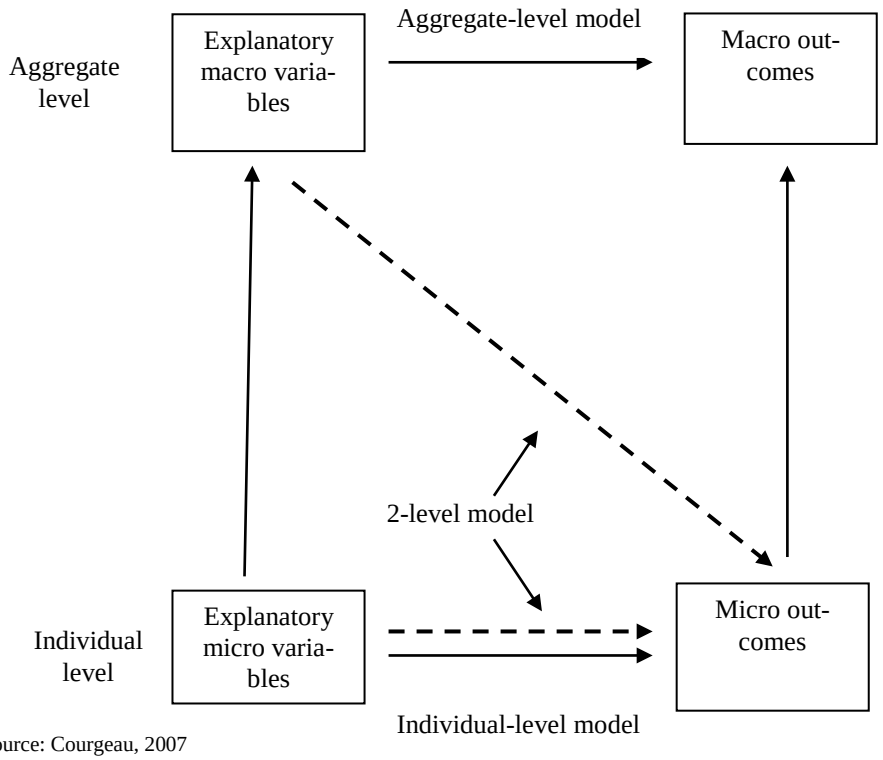
In our view, the solution to this double problem lies in multilevel analysis. It aims to introduce into a single model different aggregation levels. In addition to the individual random parameter, multilevel models include random parameters for the groups at different levels identified in the analysis. The basic assumption is that these randoms are normally distributed, so that the analysis will focus only on their variances and covariances, but may introduce individual or group characteristics at different levels.

10

Multilevel analysis no longer focuses on the group, as in the analysis on aggregate data, or on the individual, as in the event-history approach. Instead, it incorporates the individual into a broader set of levels. It thus resolves the antagonism between holism and methodological individualism. As Franck noted in 1995:

> Once we have admitted the metaphysical or metadisciplinary concept of hierarchy, it no longer makes sense to choose between holism and atomism, and—as regards the social sciences—between holism and individualism.

Figure 1 summarizes the connections between two levels, depending on whether we study them separately as in event-history and sequential models, or jointly as in a multilevel model.



Source: Courgeau, 2007

Figure 1: Connection between different levels

This approach requires new types of surveys to capture and define the various levels to examine (Courgeau, 2007). It has been used in education science, demography, epidemiology, economics, ecology, and other disciplines to identify a multilevel structure of society. In particular, multilevel event history permits a synthesis of the two approaches (Courgeau, 2003). They are semi-parametric as before,

but in some cases they can take non-parametric forms, as in multilevel factor models (Goldstein, 2003).

It privileges also the use of the Bayesian paradigm in order to deal appropriately with nested or clustered data (Draper, 2008). However some other models use the frequentist paradigm. As discussed by Greenland (2000) the multilevel approach permits to unify these two paradigms, leading to empirical Bayes estimation encompassing the two approaches. About such a convergence the interested reader may read the recent book by Schweder and Hjort (2016) on statistical inference for epistemic probability understood as confidence distributions.

In its turn, the new approach encountered certain problems, which we shall now examine.

The first problem lies in the fact that it frequently uses, as group characteristics, mean values of each member's individual characteristics or even variances or covariances. In fact there is a need to know more detailed characteristics of the aims and rules prevailing in a group and how to define them in order to explain a collective action. What are the mechanisms of social influence which permit the emergence of a collectively owned social capital in different social contexts, which "is more than the sum of the various kinds of relationship that we entertain" (Adler and Kwon, 2002)?

The second problem is that "independence among the individuals derives solely for common group membership" (Wang et al., 2013). In fact, the groups are generally more complex. For example a family, generally taken as a simple group, is in fact a more complex one where parents and children play very different and even conflicting roles. This dissymmetry of roles partly undermines the value of the family for multilevel analysis, in which we are looking for what unites group members rather than what divides them. Here again, we should take into account the interactions between group members and their changes over time in order to fully incorporate their social structure. This task will require new observation tools and new analytical methods. We will see in *2.4* how a multilevel network approach permits to avoid this problem.

The third problem lies on the difficulty to define valid groups and to use existing geographic or administrative groupings, which have little to do with their inhabitants' behaviors. Only the increase of observed existing networks in future more detailed surveys, such as those included in the *Stanford Large Network Data Set Collection*, will permit to avoid these unsatisfactory groupings.

Last, while multilevel analysis enables us to incorporate a number of known aggregation levels that constitute a society, it continues to focus only on one of these levels—an event, an individual or a group. So that this "approach assumes that links between groups are non existent" (Wang et al. 2013). Contrary to this idea it is important to take the analysis further by trying to identify the interactions that necessarily exist between the various levels. As Robert Franck wrote in 1995: "the point now is to determine how the different stages or levels connect, from top to bottom and from bottom to top". We must therefore develop a deeper study of the behaviors specific to each level but, above all, we must to try to connect the levels together in both directions.

12

We will see now how social networks permit to solve a major part of these problems.

2.4 From a network to a multilevel network approach

While earlier examples exist, research on social networks effectively began with the work of the sociologists Moreno and Jennings in the 1930's, particularly with a paper they wrote in 1938 in which they used the term 'network theory' and proposed statistics of social configurations. However almost up to the 1970's, if research teams in various social sciences worked on network analysis, no integrated cumulative effort resulted (Freeman, 2004). During the 1970s and 1980s social networks take off as a field, under the development of structural models inspired by White et al. (1976) and Freeman (1989), which examine the interdependent relationships between actors and the similar relationships between the positions of these actors in the different social networks.

This approach rests on robust mathematical foundations, which are however very different from the previous ones, as the assumption of independence of observations on individuals no more holds: network analysis argues that units are no more acting independently, but influence each other. The use of graph theory and matrix analysis is important in this field. These methods are described in detail in the book of the sociologists Wasserman and Faust (1994) as well as in the paper by the physicist Newman (2003). Many sciences, not only social, adopted this approach: information science, computer science, management, communication, engineering, economics, psychology, political science, public health, medicine, physics, sociology, geography, demography, etc.

More recently a multilevel network analysis was developed and permitted to make the link with multilevel analysis (Lazega and Snijders, 2016). While network theory is generally analyzing one given level, this approach is looking not only at the networks existing within different levels but also at the links existing between these levels. It leads to important extensions of existing models representing social structure, with networks as the dependent variable. A first kind of models tries to "reveal the interdependencies among the micro-, macro-, and meso-level networks", the meso level being here "defined between nodes of two adjacent models" (Wang et al., 2013). They generalize graph models for multiple-networks. A second kind of models "accommodate multiple partially exchangeable networks, as well as treatment effects and other covariate effects on network structure" (Sweet et al., 2013). They are often called hierarchical network models and are a generalization of multilevel models. A third kind of models "is to partition the units at all levels into groups by taking all available information into account and determining the ties among these groups" (Žiberna, 2014). It is a generalization of classical blockmodeling developed for single relations.

As for the multilevel approach, many of these models use Bayesian estimators, which have algorithmic advantages particularly for non-nested data structures, and Markov Chain Monte Carlo (MCMC) algorithms. As for multilevel models they

use simultaneously the frequentist paradigm and this common use may lead to more general empirical Bayes estimators (Greenland, 2000).

This approach requires surveys capable of capturing the different networks simultaneously. For example, the demographic survey on networks of relationships (Courgeau, 1972) captured the family, occupational, friendly, and community relationships of individuals living in a rural area. A network analysis of this survey by Forsé (1979) permitted to construct, from a complete diagram of acquaintance networks, "sociability" groups distinguished by social and demographic characteristics. Many other examples on more restricted networks include a biomedical research network, an isolated monastery, and so on (White et al., 1976), or on large scale networks such as those given in the *Stanford Large Network Data Set Collection* containing social, citation, collaboration, internet networks, etc. (see their use in Leskovec et al, 2009).

What new problems will this approach encounter?

A first problem lies in the surveys or on the existing data collections used in order to get the ties between individuals or between levels. They will never be exhaustive and, as many possibilities exist for their limitation, this may lead to important implications for their study. Very often in surveys, only a limited number of ties is asked and this number may vary from one survey to the other. There is also ambiguity about the qualification of these ties: the term "best friends" may have a different signification than "more frequently met" or 'more trusted" person. If a survey may ask for different kind of networks (family, friends, people at work, etc.), generally an existing data collection, like people on Facebook, will not permit this distinction. Even some persons may report more connections with popular, attractive or powerful persons than there are in reality.

A second problem lies in the fact that network clusters are generally created by the researcher rather than pre-existing to him. The way used to create them need many decisions that are difficult to pose in an entire scientific way. As Žiberna (2014) said:

> In conceptual terms, the main disadvantages are that there are no clear guidelines concerning what are the appropriate restrictions for ties between levels and what are appropriate weights for different parts of multi-relational networks, that is for level specific one-mode networks and for the two-mode networks.

Even if this citation is more linked to his blockmodeling approach, it is also true for a more general multilevel network approach. In every case decisions must be made by the researcher how to include or exclude people, merge or divide network clusters, etc. But this also permits the statistical analysis when these networks are clearly defined.

A third problem lies in the difficulty to introduce individual or network characteristics in the study of these networks. Only the use of hierarchical network models permit to introduce them. But, even in this case, there are few data sets which give measures of covariate effects on network structure (Sweet et al., 2013). These covariates may be individual, network, tie-specific, or a combination of the three.

14

A fourth problem lies in the introduction of time in these studies. Again, very few surveys permit to observe the changes occurring to networks through time. Some multi-wave surveys give at different times the structure of a network. Lazega et al. (2011) used a three wave survey in order to show that the structure of an organization remains the same regardless of the turnover of the members. However there is a need of more detailed longitudinal observations at multiple levels of analysis and of new methods in order to study the organizational mobility and relational turnover implied by the introduction if time in multilevel networks.

We can conclude this examination of the different problems and challenges encountered by network and multilevel network analysis, by the conclusion given by Lazega and Snijders in their 2016 book:

> Among the most difficult (challenges), we find combining network dynamics and multilevel analysis by providing statistical approaches to how changes at each level of collective agency drive the evolution of changes at other levels of collective agency. In all these domains, much remains to be done.

So that we can think that these problems are more a challenge for this approach than unsolvable ones.

We will now turn to the last approach presented in this paper.

2.5 An approach by agent-based decisions

Individual- or agent-based models[2] constitute an approach that differs much more from the previous ones. These models are derived from the analyses of simulation by the mathematicians Von Neumann and Ulam and the physicist Metropolis (1949). The economist Schelling (1971) suggested their use to study segregation processes and in 1972 the ecologists Botkin et al. proposed a computer model in order to predict the evolution of forest growth. During the 90's these models spread to different social sciences, taking often care not to consider each science separately but on the contrary to view them as a whole incorporating all the various social processes—demographic, economic, sociological, political, and so on. They are now largely used in many domains.

Rather than modeling specific data, this approach models theoretical ideas and is based on computer simulation. Its aim is to understand how the behavior of biological, social, or more complex systems arises from the characteristics of the individuals or more general agents making up these systems. As Billari et al. (2003) said:

---

[2] In general the ecologists prefer to speak of individual-based models, while the social scientists prefer the term agent-based, but the two denominations recover quite the same approach. We will use here the denomination of agent-based models usual for the social sciences.

15

> Different to the approach of experimental economics and other fields of behavioral science that aim to understand why specific rules are applied by humans, agent-based computational models pre-suppose rules of behavior and verify whether these micro based rules can explain macroscopic regularities.

So that this approach is bottom-up, with population-level behavior emerging from rules of behavior of autonomous individuals. These models are described in a number of books, as for example Epstein (2007) in social science or Railsback and Grimm (2012) in ecology. Many other natural and social sciences adopted it, including physics, ecology, archaeology, demography, sociology, computer science, economics, epidemiology, political science, etc.

This agent-based approach eliminates the need of empirical data on personal or social characteristics in order to explain a phenomenon as it is based on simple rules of decision followed by individuals, which can explain some real-world phenomenon. As Burch (2003) said:

> A model explains some real-world phenomenon if a) the model is appropriate to the real-world system …, and b) if the model logically implies the phenomenon, in other words, if the phenomenon follows logically from the model as specified to fit a particular part of the real world.

Such a theoretical model cannot be validated in the same way than an empirical model, as the four previously presented approaches. About this approach Franck (2002) said: "… one has ceased to credit deduction with the power of explaining phenomena. Explaining phenomena means discovering principles which are implied by the phenomena." As it focuses on the mechanisms which drive the action of individuals or agents, it will simulate the evolution of such a population from simple rules of behavior. So that it may use game theory, complex system theory, emergence, evolutionary programming and, in order to introduce randomness, Monte Carlo methods. It may also use survey data, not in order to explain the studied phenomenon, but only to verify if the parameters used in the simulation lead to a similar observed behavior as in the survey. For example Heiland (2003) used an agent-based model in order to recover the observed distribution of migrants across different West German states and over a period of 9 years (1989-1997) from an Eastern state (Sachsen). With few theoretical assumptions about the decision to migrate the simulations indicate that heterogeneity in mobility can explain the observed decline in migration.

Again these agent-based models raised some new problems.

The first problem is that these models "are intended to represent the import and impact of individual actions on the macro-level patterns observed in a complex system" (Courgeau et al., 2016). This implies that an emergent phenomenon at the aggregate level can be entirely explained by individual behavior. However Holland (2012) said about agent-based models that they include "little provision for agent conglomerates that provide building blocks and behavior at higher level of organization". In fact a study by a multilevel model on the effects of an individual characteristic (being farmer) and the corresponding aggregate one (the proportion of farmers living in an area) on the probability of internal migration in Norway

(Courgeau, 2007) shows that these effects are opposite ones. It seems then difficult to explain a macro-characteristic acting positively by a micro-characteristic acting negatively. Indeed, micro-level rules find hardly a link with aggregate-level rules, and I think that aggregate-level rules cannot be modeled with a micro-approach, since they transcend the behaviors of the component agents.

A second problem is that this approach is mainly bottom-up. As we have already previously seen for multilevel network models it is important to consider simultaneously a top-down process from higher-level properties to lower-level entities. More precisely we will have to speak about a Micro-Macro link which "is the loop process by which behaviour at the individual level generates higher-level structures (bottom-up process), which feedback the lower level (top-down), sometimes reinforcing the producing either directly or indirectly" (Conte et al., 2012). The bottom-up approach of agent-based models is unable to take into account such a Micro-Macro link.

A third problem lies on the validation of a given agent-based model. Such an approach is an attempt for imitation of human behavior using some well chosen generative mechanisms to produce it. It may be judged as successful when it leads to a correct reproduction of the structural characteristics of this behavior. The way to ascertain this judgment is however very far from usual tests used to verify the validity of the effects of different characteristics in the previous approaches. Such a test which can be made in natural science is less evident in social science. As Küppers and Lenhard (2005) said:

> The essential point is that (often) in the natural sciences one has a general theory about the objects and simulation models are used as instruments to generate data and to make predictions about the behaviour of these objects. On the other side, agent-based models are instruments to explore the theoretical structure of the data.

In order to see if such an exploration had been successful, we need to consider different aspects. First, how to test that there are no other models able to explain better the observed phenomenon? Often the researcher tries different kind of models in order to permit to choose the one which give the better accord with empirical data. But this does not solve the problem, as there is an infinity of models which may predict the same empirical result as well or better. The agent-based approach gives no way to avoid this problem. Second, how to test that the chosen model gives a good fit to the observed data? Unfortunately, there are no clearly defined procedures for testing the fit of the simulation models, like goodness of fit procedures or tests of significance for the previous approaches. We can conclude that there are no clear verification and validation procedures for agent-based models in population science.

In consequence we will have to see in the next section how to try to overcome these problems, as well as those encountered with the four previous approaches.

3. Towards a synthesis

As we have already presented and criticized five different main approaches used nowadays in social sciences, we will have now to see if we can give a more synthetic view of them.

Let us first consider the two main concepts without which no population science would be possible.

The first one is the creation of an abstract fictious individual, whom we can call a *statistical individual* as distinct from an *observed individual*. As Aristotle (330 BC) said: "individual cases are so infinitely various that no systematic knowledge of them is possible", Graunt (1662) was the first to introduce the possibility of a population science letting aside the observed individual and using statistics on a few number of characteristics, leading to a statistical individual. As Courgeau wrote in 2012:

> Under this scenario, two observed individuals, with identical characteristics, will certainly have different chances of experiencing a given event, for they will have an infinity of other characteristics that can influence the outcome. By contrast, two statistical individuals, seen as units of a repeated random draw, subjected to the same sampling conditions and possessing the same characteristics, will have the same probability of experiencing the event.

The essential assumption permitting to use the theory of probability in this case is that of *exchangeability*[3] (de Finetti, 1937): *n* trials will be said to be exchangeable if the joint probability distribution is invariant for all permutations of the *n* units. We will use it here for the residuals given the explanatory characteristics measured on these individuals.

The second concept is the notion of a *statistical network*, different from the observed ones: it appeared more recently, for example with the work of Coleman in 1958. While *observed networks* may be as diverse as the infinite kind of ties existing between observed individuals, *statistical networks* may be more precisely defined with the use of statistics on ties and the choice of criteria to circumscribe them. Again the essential assumption permitting the use of the theory of probability is that, given the explanatory characteristics introduced at each level, the residuals are assumed to be exchangeable.

It is interesting here to compare these two concepts with the contexts proposed by Billari (2015) to explain population change: the micro- and the macro-level contexts. In fact he clearly recognized at the basis of micro-level context the abstract concept of *statistical individual*, the same that we propose here. However for macro-level context he only proposes to see how "population patterns reemerge from action and interaction of individuals", without recognizing the abstract concept at the basis of this interaction: the *statistical network*, which permits to flesh out this macro-analysis. For example we have already seen how multilevel analysis permits to reconcile the macro- and micro-level results.

[3] In this first paper on this topic de Finetti called it *equivalence*.

18

Once these two main concepts defined, we can see that the study of event duration and the study of event sequences are directly connected to the same concept of *statistical individual*. Even if their approach of this individual is different, as we have already seen, they can be considered as two complementary ways to study him. In addition some more recent papers, as those presented in this conference by Studer et al. (2016) or Rossignon et al. (2016), combines the advantages of the two approaches modeling "the relationships between time varying covariates and trajectories specified as processes outcomes that unfold over time". The definition given by Courgeau and Lelièvre (1997) of event history analysis appears to be also valid for sequence analysis: "Throughout his or her life an individual follows a complex itinerary, which at a given moment depends on the life course to date and on the information acquired in the past". The itinerary is followed event after event, in the first analysis, and with more complex sequences of events in the second one.

Similarly we can see that the contextual, multilevel and network multilevel approaches are simultaneously connected to the same concept of *statistical network*. They appear as complementary in its study. We can say that contextual and multilevel analysis focuses on attributes while network multilevel analysis focuses on relations, combining the different levels.

It may also be interesting to see that contextual and multilevel analysis may be seen as complementary of event history analysis, introducing the effects of network membership on individual behavior. Similarly network multilevel analysis may also be seen as complementary to sequence analysis. This proximity may explain why Cornwell (2015) tries to introduce network methods in sequence analysis: however sequence methods remain at the statistical individual level, while multilevel networks methods introduce statistical networks.

The different problems encountered when using one of these four approaches may largely disappear when considering simultaneously the statistical individual and the statistical network under a more general *biographical multilevel network analysis*. As we already said such an approach is able to avoid the risks of atomistic or of ecological fallacy through the use of a synthesis of holism and individualism. It may also avoid the problems linked to the choice between Bayesian or frequentist probability through the use of a more general compromise on confidence distributions (Schweder and Hjort, 2016), which opens to a better statistical inference. It permits to answer to some problems posed by unobserved heterogeneity, while introducing networks which permit to have a better understanding of human behavior. We can also think that a number of problems encountered by sequence analysis (metric used, cluster analysis and artifacts) may be solved by undertaking more complex surveys on social networks, which may permit to replace theoretical clusters by real networks of individuals linked together by existing social forces. Similarly the main problems encountered by multilevel analysis may largely be solved by multilevel network analysis, such as: the use of a Multilevel Social Influence (MSI) model (Agneessens and Koskinen, 2016) to explain the emergence of a social capital; the use of Exponential Random Graph Models (ERGM) to show that within-level network structure are interdependent with network structures of other levels (Wang et al., 2016); etc. Last we think that the problems re-

cently posed by multilevel network analysis are more a challenge for future research in this field, than unsolvable problems. For example such an analysis will reach its full potential when longitudinal observations at multiple levels of analysis will be available, by providing a combination of an event history of networks with a multilevel analysis (Lazega and Snijders, 2016).

The situation is more complex for agent-based approach. While it apparently resembles event-history approach in its focus on individual behavior alone, it seeks however to explain collective behavior with the aid of individual behavior. This gives it some affinity with the multilevel network approach. The main question is: how to generate the macroscopic regularity from the bottom-up, using simple local rules? The difficulties encountered with such an approach are clearly described in Conte et al. (2012):

> First, how to find out the simple local rules? How to avoid *ad hoc* and arbitrary explanations? As already observed[4], one criterion has often been used, i.e., choose the conditions that are sufficient to generate a given effect. However, this leads to a great deal of alternative options, all of which are to some extent arbitrary.

As we have already shown, in social science we cannot obtain the macro-level patterns by simply aggregating the micro-level outcomes, so that local rules are not sufficient to explain a complex social behavior. It is then necessary to introduce theories of decision making to get more valuable models: however the number of options for modeling decision making is almost infinite (Klabunde and Willekens, 2016). As the choice of a decision theory is driven by the researcher background, an economist, a demographer, a geographer, a psychologist, etc., may reach quite different results for the same studied phenomenon.
In our view, further work is needed to go over these contradictions and place agent-based analysis in a broader setting and a more explicit theory-founded model.

4 Conclusion

By restricting ourselves to defining a scientific method solely by its methods, we condemn ourselves to taking a partial view of the core scientific approach. We need to set up a more robust research program for demography and, more generally, the social sciences—a program that converges with the now well established program of the physical and biological sciences. The source for this program lies in Bacon's work (1620):

> There are and can be only two ways of searching into and discovering truth. The one flies from the senses and particulars to the most general axioms, and from these principles, the truth of which it takes for settled and immovable, proceeds to judgment and to the discovery

[4] See Conte (2009) p. 29.

20

> of middle axioms. And this way is now in fashion. The other derives from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms last of all. This is the true way, but as yet untried.

Bacon calls the second approach *induction*, not in the meaning later given to the term by the empiricist tradition of Hume and Popper—i.e., the generalization of observations—but in the sense of the search for the structure of observed phenomena. That is how Galileo, Newton, Graunt, Einstein, Darwin, and others developed their approach to the study of phenomena—whether physical or social.

It is important for the social sciences to start with the observation and measurement of facts, for this measurement, far from being secondary, makes it possible to assess the "potentialities" of a social fact (Courgeau, 2013). Next, instead of relying on often arbitrary hypotheses, like in agent-based models, the modeling of observed phenomena should follow the method recommended by Bacon by analyzing the interactions between the networks created by people and seeking their structure (Franck, 2002; Courgeau et al., 2016).

Even if one can think that each individual has an unlimited and unknowable number of characteristics with his own freedom of choice, social science has to see that he is born in a given society with its rules and laws, which restrain his freedom, that he is submitted to biological laws, which are the same for all humans. So that a social science can exist which takes into account only a limited number of characters and which is based on a number of concepts without which the properties of these characters would be inconceivable or impossible (Franck, 2002).

A final point: We have often viewed the social sciences here as a whole to which certain approaches applied and not others. We must now consider that it is not by erasing the boundaries between disciplines that we can improve our knowledge (Franck, 1999). The boundaries are real, for each discipline endeavors to analyze different properties of human societies. However, we think that it is possible to construct a new formal object that can explain certain properties of human societies—an object that encompasses existing disciplines and allows their synthesis.

References

Aalen, O.O. (1975). *Statistical inference for a family of counting processes*. PhD thesis, Berkeley: University of California.

Aalen, O.O., Borgan Ø, Keiding, N., Thorman, J. (1980). Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics*, 7, 161-171.

Aalen, O.O., Borgan, Ø, Gjessing, H, K (2008). *Survival and event history analysis. A process point of view*. New York: Springer.

Abbott, A. (1983). Sequences of social events: concepts and methods for the analysis of order in social processes. *Historical Methods*, 16 (4), 129-147.

Abbott, A. (1984). Event sequence and event duration: colligation and measurement. *Historical methods*, 17 (4), 192-204.

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annual Review of sociology*, 21, 93-113.

Abbott, A., Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: review and prospect. *Sociological Methods and Research*, 29, 3-33.

Agneesens, F., Koskinen, J. (2016). Modeling individual outcomes using a Multilevel Social Influence (MSI) model: individual versus team effects of trust on job satisfaction in an organisational context. In T. Lazega, T. Snijders (eds) (2016). *Multilevel network analysis for the social sciences*, Methodos Series, vol.12, pages 81-105, Dordrecht / Boston / London: Springer

Alder, P.S., Kwon, S.-W. (2002). Social capital: prospects for a new concept. *Academy of Management*, 27 (1), 17-40.

Andersen, P.K., Borgan, Ø, Gill, R.D., Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.

Aristotle (around 330 B.C.). *Rhetoric*. The Internet Classic Archive: translated by Roberts, W.R.

Bacon, F. (1620). *Novum Organum*, London: J. Bill.

Billari, F. (2015). Integrating macro- and micro-level approaches in the explanation of population change, *Population Studies*, 69 (S1), S11-S20.

Billari, F., Prskawetz, A., ed. (2003). *Agent-based computational demography: using simulation to improve our understanding of demographic behaviour*. Heidelberg, New York: Physica-Verlag.

Bison, I. (2009). OM matters: the interaction effects between *Indel* and substitution costs. *Methodological Innovations Online*, 4 '2), 53-67.

Bison, I. (2014). Sequences as network: an attempt to apply network analysis to sequence analysis. In P. Blanchard, F. Bühlman, J.-A. Gauthier (eds), *Advances in sequence analysis: theory, method, applications*, pages 231-248, Dordrecht: Springer.

Blanchard, P., Bühlman, F., Gauthier, J.-A. (eds) (2014). *Advances in sequence analysis: theory, method, applications*. Dordrecht: Springer.

Bolano, D. (2014). Hidden Markov models: an approach to sequence analysis in population studies. *Annual meeting of the Population Association of America*, presented in Poster Session, paa2014.princeton.edu/papers/141879, 22 p.

Botkin, D.B., Janak, J.F., Wallis, J.R. (1972). Some ecological consequences of a computer model of forest growth. *The Journal of ecology*, 60, 849-872.

Breiger, R.L. (1974). The duality of persons and groups. *Social Forces*, 53, 181-190.

Bretagnole, J. & Huber-Carol, C. (1988), Effects of omitting covariates in Cox's model for survival data, *Scandinavian Journal of Statistics*, 15, 125-138.

Burch, T.K. (2003). Data, models, theory and reality: the structure of demographic knowledge. In F.C. Billari, A. Prskawetz (eds), *.Agent-based computational demography*, pages 19-40, Heidelberg New York: Physica Verlag.

Byrne, D., Uprichard E. (2012). Introduction; (useful) hey texts. In D. Byrne, E. Uprichard (eds), *Cluster analysis*, vol. 2, pages vii-xii, London: Sage Publications Ltd.

Coleman, J.S. (1958). Relational analysis: the study of social organizations with survey methods. *Human Organization*, 17, 28-36.

Conte, R. (2009). From simulation to theory (and backward). In F. Squazzoni (ed), *Epistemological aspects of computer simulation in the social sciences*, pages 29-47, Berlin Heidelberg: Springer.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214, 325-346.

Cornwell, B. (2015). *Social sequence analysis*. New York: Cambridge University Press.

Courgeau, D. (1972). Les réseaux de relations entre personnes. Etude d'un milieu rural. *Population*, 27 (4-5), 641-684.

22

Courgeau, D. (1999). L'enquête « Triple biographie : familiale, professionnelle et migratoire ». In Groupe de réflexion sur l'approche biographique, *Biographies d'enquêtes*, pages 59-74, Paris: INED.

Courgeau, D., ed. (2003). *Methodology and epistemology of multilevel analysis: approaches from different social sciences.* Methodos Series, vol. 2, Dordrecht / Boston / London: Kluwer Academic Publishers.

Courgeau, D. (2007). *Multilevel synthesis: from the group to the individual*. Series on demographic methods and population analysis, vol. 18, Dordrecht: Springer.

Courgeau, D. (2012). *Probability and social science. Methodological relationships between the two approaches*. Methodos Series, vol. 10, Dordrecht / Boston / London: Kluwer Academic Publishers.

Courgeau, D. (2013). La mesure dans les sciences de la population. *Cahiers Philosophiques*, 135 (4), 51-74.

Courgeau, D., Bijak, J., Franck, R., Silverman, E. (2016). Model-based demography: towards a research agenda. In A. Grow, J. Van Bavel, (eds), *Agent-based modelling in population studies*, Series on demographic methods and population analysis, vol. 41, Dordrecht: Springer.

Courgeau, D., Lelièvre, E. (1992). *Event history analysis in demography*. Oxford: Clarendon Press.

Courgeau, D., Lelièvre, E. (1997). Changing paradigm in demography. *Population, An English Selection*, 9, 1-10.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, 34 (2), 187-220.

Cox, D.R. & Oakes, D. (1984). *Analysis of survival data*. London, New York: Chapman and Hall.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7 (1), 1-68.

Degenne, A., Forsé, M. (1999). *Introducing social networks*. London: Sage.

Doob, J.L. (1953). *Stochastic processes*. New York: Wiley.

Durkheim, E. (1895). *Les règles de la méthode sociologique*. Paris: Alcan.

Epstein, J. (2007). *Generative social science: studies in agent-based computational modelling*. Princeton: Princeton University Press.

Everitt, B.S. (1979). Unresolved problems in cluster analysis. *Biometrics*, 35, 169-181.

Fararo, T.J., Doreian, P. (1984). Tripartite structural analysis: generalising the Beiger-Wilson formalism. *Social Networks*, 6, 141-175.

Forsé, M. (1979). Les réseaux de sociabilité dans un village. *Population*, 36 (6), 1141-1162.

Franck, R. (1995). Mosaïques, machines, organismes et sociétés. *Revue Philosophique de Louvain*, 93 (1-2), 67-81.

Franck, R. (1999). La pluralité des disciplines, l'unité du savoir, et les connaissances ordinaires. *Sociologie et Sociétés*, 1, 129-142.

Franck, R., Ed. (2002). *The explanatory power of models: bridging the gap between empirical and theoretical research in the social sciences.* Boston / Dordrecht / London: Kluwer Academic Publishers.

Freeman, L.C. (1989). Social networks and the structure experiment. In L. C. Freeman, D. R. White and A. K. Romney (eds), *Research Methods in Social Network Analysis*, pages 11-40, Fairfax: George Mason University Press.

Freeman, L.C. (2004). *The development of social network analysis: a study in the sociology of science*. Vancouver: Booksurge Publishing.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least-squares. *Biometrika*, 73, 43-56.

Goldstein, H. (2003). *Multilevel statistical models.* London: Edward Arnold.

Granger, G.-G. (1994). *Formes, opérations, objets*. Paris: Librairie Philosophique Vrin.

Graunt, J. (1662). *Natural and political observations mentioned in a following index, and made upon the bills of mortality*. London; Tho. Roycroft.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, 39, 156-167.

Heiland, F. (2003). The collapse of the Berlin wall: simulating state-level East to West German migration patterns. In F.C. Billari, A. Prskawetz (eds), *Agent-based computational demography: using simulation to improve our understanding of demographic behaviour*, pages 73-96, Heidelberg, New York: Physica-Verlag,.

Henderson, D.F., Kempthorne, O., Searle, S.R., Von Krosigk, C.M. (1959). The estimation of environmental and genetics trends from records subject to culling. *Biometrics*, 15, 192-218.

Holland, J.H. (2012). *Signals and boundaries. Building blocks for complex adaptive systems*. Cambridge London: The MIT Press.

Holland, P.W., Leinhardt, S. (1976). Local structure in social network. *Sociological Methodology* 7, 1-45.

Ibrahim, J.G., Chen, M.-H., Sinha, D. (2001). *Bayesian Survival analysis*. New York: Springer-Verlag.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon Press.

Kalbfleisch, J.D., Prentice, R.L. (1980). *The statistical analysis of failure time data*. New York, Chichester, Brisbane, Toronto: John Wiley and Sons.

Klabunde, A., Willekens, F. (2016). Decision-making in agent-based models of migration: state of the art and challenges. *European Journal of Population*, 32, 73-97.

Küppers, G., Lenhard, J. (2005). Validation of simulation: patterns in the social and natural science. *Journal of Artificial Societies and Social Simulation*, 8 (4), 1-13.

Lazarsfeld, P.F., Menzel, H. (1961). On the relation between individual and collective properties. In R. Boudon (ed.), *Complex organizations: a sociological reader*, pages 172-189, Chicago: University of Chicago Press.

Lazega, T., Sapulete, S., Mounier, L. (2011). Structural stability regardless of membership turnover? The added value of blockmodelling in the analysis of network evolution. *Quality & Quantity*, 45, 129-144.

Lazega, T., Snijders, T., eds., (2016). *Multilevel network analysis for the social sciences*. Methodos Series, vol.12, Dordrecht / Boston / London: Springer.

Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M. (2009). Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6 (1), 29-123.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, 707-710.

Levitt, M. (1969).Detailed molecular model for transfer ribonucleic acid. *Nature*, 224 (5221), 759-763.

Liu, J.S., Logvinenko, T. (2003). Bayesian methods in biological sequence analysis. In D.J. Balding, M. Bishop, C. Cannings (eds), *Handbook of statistical genetics*, 3rd Edition, pages 67-96, Chichester: John Wiley & Sons, Ltd.

Loriaux, M. (1989). L'analyse contextuelle : renouveau théorique ou impasse méthodologique? In J. Duchêne, G. Wunsch, E. Vilquin (eds), *Explanation in the social sciences. The search for causes in demography*, *Chaire Quetelet '87*, pp. 333-368, Bruxelles: Editions Ciaco.

Mason, W.M., Wong G.W., and Entwistle B. (1983), Contextual analysis through the multilevel linear model, in *Sociological Methodology 1983-1984*, Leinhart S. ed., San Francisco: Jossey-Bass, pp. 72-103.

Metropolis, N., Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44 (247), 335-341.

Mill, J.S. (1843). *A system of logic, ratiocinate and inductive, being a connected view of the principles of evidence, and the methods of scientific investigation*, vol. I, London: Harrison.

Moreno, J.L., Jennings, H.H. (1938). Statistics of social configurations. *Sociometry*, 1 (3/4), 342-372.

Newman, M.E. (2003). The structure and function of complex networks. *Siam Review*, 45 (2), 167-256.

Petty, W. (1690). *Political arithmetick*. London: Robert Clavel & Hen. Mortlock.
Railsback, S.F., Grimm, V. (2012). *Agent-based and individual-based modelling*. Princeton: Princeton University Press.
Robette, N., Bry, X. (2012). Harpoon or bait: a comparison of various metrics to fish for sequence patterns. *Bulletin de Méthodologie Sociologique*, 116 (1), 15-24.
Robette, N., Lelièvre, E., Bry, X. (2012). La transmission des trajectoires d'activité : telles mères, telles filles. In C. Bonvalet, E. Lelièvre (eds), *De la famille à l'entourage*, pages 395-418, Ined, Paris.
Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.
Rossignon, F., Studer, M., Gauthier, J.-A., Le Goff, J.-M. (2016). Childhood family structure and home-leaving. A combination of survival and sequence analysis. *LaCOSA II Conference*, 11 pages, Lausanne.
Rozenblat, C., Melançon, G. (2013). *Methods for multivariate analysis and visualization of geographical networks*. Methodos Series, vol. 11, Dordrecht / Boston / London: Springer.
Schelling, T. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1, 143-171.
Schweder, T., Hjort, N.L. (2016). *Confidence, likelihood, probability: statistical inference with confidence distributions*. Cambridge: Cambridge University Press.
Studer, M., Struffolino, E., Fasang, A. (2016). A new tool for old questions: the sequence-analysis multistate model to study relationships between time-varying covariates and trajectories. *LaCOSA II Conference*, 31 pages, Lausanne.
Sweet, T.M., Thomas, A.C., Junker, B.W. (2013). Hierarchical network models for education research: hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, 38, 295-318.
Trussell, J. (1992). Introduction. In J. Trussell, R. Hankinson, Tilton, J. (eds), *Demographic applications of event history analysis*, pages 1-7, Oxford: Clarendon Press.
Trussell, J. Richards, T. (1985), Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure, in N. Tuma (ed.), *Sociological Methodology*, pages 242-276. Jossey-Bass, San Francisco (CA).
Tryon, R. (1939). *Cluster analysis*.Ann Arbor:Edwards Brothers
Vaupel, J.W., Manton, K.G., Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439-454.
Ville, J.A. (1939). *Etude critique de la notion de collectif*. Paris: Gauthier-Villars.
Wang; P., Robins, G., Pattison, P. Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, 35 (1), 96-115.
Wang, P., Robins, G., Matous, P. (2016). Multilevel network analysis using ERGM and its extensions. In T. Lazega, T. Snijders (eds) (2016). *Multilevel network analysis for the social sciences*, Methodos Series, vol.12, pages 125-143, Dordrecht / Boston / London: Springer.
Wasserman, S., Faust, K. (1994). *Social network analysis: methods, and applications*. Cambridge: Cambridge University Press.
Wasserman, S., Iacobucci, D. (1991). Statistical modelling of one-mode and two-mode networks: simultaneous analysis of graphs and bipartite graphs. *British Journal of Mathematical and Statistical Psychology*. 44, 13-44.
White, H.C., Boorman, S.A., Breiger, R.L. (1976). Social structure from multiple networks. I Blockmodels of roles and positions. *American Journal of Sociology*, 81 (4), 730-780.
Žiberna, A. (2014). Blockmodeling of multilevel networks. *Social networks*, 39 (1), 46-61.