Do Different Approaches in Population Science Lead to Divergent or Convergent Models?

Daniel Courgeau

Do Different Approaches in Population Science Lead to Divergent or Convergent Models?

Daniel Courgeau

1 Introduction

Since its introduction by Graunt in 1662, the scientific study of population, initially called *political arithmetick*, has become possible not only for demography but also for epidemiology, political economy, and other fields. For more than 200 years, researchers adopted a *cross-sectional approach* in which social facts in a given period exist independently of the individuals who experience them, and can be explained by various characteristics of the society of that period. After the end of World War II, researchers examined social facts from a new angle, introducing individuals' life experience. This *longitudinal approach* holds that the occurrence of a given event, during the lifetime of a birth cohort, can be studied in a population that maintains all its characteristics as long as the phenomenon persists. However, this condition was too restrictive, triggering the development of new approaches that we shall discuss in greater detail here, with an emphasis on the scope for convergence or divergence.

From the comparison of these new approaches, we shall try to identify the conditions that would allow a synthesis of these approaches by means of a Baconian inductive analysis. This induction method consists in discovering the principles of a social process by experiment and observation. It is therefore based on the fact that, without these principles, the observed properties would be different. It will enable us to draw a conclusion.

D. Courgeau (✉)
Institut National d'Etudes Démographiques (INED), Paris, France
e-mail: daniel.courgeau@wanadoo.fr

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,
Life Course Research and Social Policies 10,
https://doi.org/10.1007/978-3-319-95420-2_2

2 Different Approaches

We shall describe and focus our discussion on four major approaches based on: duration between events, sequences, multiple levels, and networks; we set aside agent-based models (i.e., based on agents' decisions) which are of a totally different type. We will try to show in what these approaches are different and in what they may converge.

2.1 An Approach Based on Duration Models

The first approach emerged in the social sciences in the early 1980s, more than 30 years after the introduction of longitudinal analysis. However, it was already used by statisticians, such as Ville in 1939 and Doob in 1953, in association with the concept of martingale. It was Cox who, in 1972, recommended the simultaneous use of life tables and regression methods, and Aalen in 1975 who proposed the use of counting process theory for the simultaneous analysis of several events that an individual may experience over time.

The principle of this approach is that "throughout his or her life, an individual follows a complex life course, which depends at any moment on the past history and the information acquired previously" (Courgeau and Lelièvre 1997, p. 5).

This "event-history" approach rests on robust mathematical and statistical foundations that make it possible to determine risk factors and to process censored observations. It has been described in many statistical works since 1980 (Kalbfleisch and Prentice 1980; Cox and Oakes 1984; Aalen et al. 2008). It can be used to analyze changes of state, however diverse, and to demonstrate the role of many individual characteristics that can change over time or during transitions. The application of these methods in demography yielded significant advances with respect to longitudinal analysis (see, for example, Courgeau and Lelièvre 1992). Many other social sciences adopted it including epidemiology, biostatistics, sociology, econometrics, actuarial sciences, and medicine.

The event-history approach eliminates the need for the highly restrictive hypotheses of longitudinal analysis while maintaining the individual point of view. Individuals can be followed for relatively long spells of their lives by means of retrospective or prospective surveys that track a large number of events. For example, the French 1981 "Triple event history" survey ("3B survey") allowed a simultaneous analysis of family-related, occupational, and migration-related events occurring from their birth to the survey date, for cohorts born between 1911 and 1936 (Courgeau 1999). As the approach makes it possible to process censored observations, the persons still economically active at the time of the survey (four-fifths of the sample) could be studied over the duration of their working careers in the same way as the persons already retired (one-fifth of the sample).

The approach relies mainly on semi-parametric methods that, while maintaining a nonparametric vision of time between events, use parameters to describe the effect of personal characteristics.

However, the event-history approach raises a set of problems to which we now turn. All the problems mentioned here are more structural than technical.

The first problem concerns unobserved heterogeneity. How does this heterogeneity affect the parameters of observed characteristics? An important result obtained by Bretagnolle and Huber-Carol (1988), but often overlooked by the models' users, provides an answer to the question. These authors show that when omitted characteristics are independent of observed characteristics, the omission does not impact the signs of the estimated parameters, but only reduces their absolute values. Therefore, if the effect of a characteristic is fully significant, the introduction of the initially unobserved characteristic will increase this effect alone. Conversely, a characteristic without significant effect may have one when the unobserved characteristics are introduced (Courgeau and Lelièvre 1992). One must be aware of this risk.

When the observed and omitted characteristics are interdependent, the situation is more complex. It may be tempting to introduce this heterogeneity in the form of a distribution of a known type, which Vaupel et al. (1979) have called *frailty*. When information on this distribution is available, its introduction is entirely legitimate. The problem is that, in most cases, we know nothing about this distribution, and it is often chosen for no other valid reason than convenience. In such circumstances, some distributions may change the sign of an estimated parameter, whereas a model without frailty avoids this drawback (Trussell and Richards 1985).

We therefore totally agree with Aalen et al. (2008, p. 425), 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.

Indeed, for the analysis of non-repetitive events, there is only one model without unobserved heterogeneity, but an infinity of models with unobserved heterogeneity (Trussell 1992). Their estimates differ, but none has a solid justification. 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. Most of the statisticians who developed the method chose an *objective probabilistic approach*. Could an *epistemic approach* to probability enable us to lift some of the constraints involved in objective probability? Space precludes a full description of these different approaches, but we can focus on the constraints linked to *statistical inference* (Courgeau 2012).

The purpose of statistical inference is to optimize the use of incomplete information in order to arrive at the best decision. Statistical inference will therefore

consist in providing the best possible analysis of a past phenomenon and the best possible 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 disciplines such as medicine or 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 for achieving a specific effect. Statistical inference leads, among other things, to testing various hypotheses about the phenomena studied.

Objectivist methods seek to verify whether a given factor does or does not affect the phenomenon studied. This brings us to the notion of statistical test, which involves treating the sample under analysis as one possible selection from an infinity of samples that we could 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 wish to conclude that the probability of the unknown parameter lying in the interval is 0.95. In fact, however, the objectivist tells 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 lie in that interval 95% of the time. As Jeffreys wrote (1939, p. 377):

> 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 objectivist statistical tests. Similarly, the use of frequentist methods—here, 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 does not allow us to factor in the uncertainty of the parameter estimations, and it will cause us to underestimate the variance of the predicted distribution.

This is why Jeffreys showed that, if we adopt an epistemic approach, a 95% confidence interval will mean precisely an interval in which the statistician can conclude that the unknown parameter will lie with a probability of 0.95. Moreover, this approach offers a more satisfactory resolution of the prediction problem than we could obtain with objective probability. 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 single value, as with the objectivist solution, but a distribution.

These advantages of an epistemic method have led a number of authors to propose it for event-history analysis (Ibrahim et al. 2001).

The final problem we would like to discuss is the risk of atomistic fallacy involved in the event-history approach (Robinson 1950). If we can take all individual characteristics into account to explain a behavior, we may overlook the context in which the behavior occurs. If, instead, we use a cross-sectional approach, we only introduce the characteristics of society to explain social facts. This aggregate approach is, by contrast, vulnerable to the risk of ecological fallacy. We can easily show that the relationships between two characteristics measured on individuals

or on proportions applied to different aggregates are generally far from identical (Courgeau 2007). In the third subsection of this section, we shall see the solution for overcoming these divergences.

2.2 An Event Sequences Approach

Sequence analysis was first introduced in computer science (Levenshtein 1966), then in molecular biology to study DNA and RNA sequences (Levitt 1969). It was imported into the social sciences by the sociologist (Abbott 1983, 1984) to study the social processes occurring in sequences over a long period.

The main principle of this approach is that social organization derives "from the regular and predictable pattern of temporal, spatial, hierarchical and other ordered phenomena that result". (Cornwell 2015, p. 24–25). So that the key assumption is that there is such a pattern and that it is socially meaningful. Its purpose appears to be quite different from the one of the previous models, and is now based in the order rather than in the duration of phenomena.

In the social sciences, however, this approach rests on less robust mathematical and statistical foundations than event-history analysis. In its most common setting, its goal is to describe complete sequences in terms of types reflecting socially significant trajectories of subjects (individuals or more general entities such as stimuli in psychology or artifacts in archeology). The approach comprises two stages. First, it tries to calculate a distance between sequences with the aid of certain operations (insertions or deletions called "indels" or substitutions) with a given cost for each operation. The most widely used metric is called Optimal Matching (OM), but we shall discuss other methods for calculating distances below. In the second stage, cluster analysis is used to detect types of sequences, grouping subjects into mutually exclusive categories. Many natural, biologic and social sciences use these methods, including computer science, biology, sociology, demography, psychology, anthropology, political science, and linguistics.

With sequence analysis, we can move from a Durkheimian search for causes (1895), to an emphasis on contexts, connections, and events—a shift about which (Abbott 1995, p. 93) says "a quiet revolution is underway in social science." The surveys tracking individuals for long spells of their lives resemble event-history surveys, with a focus on observing complete processes without censorship. Their objectives, by contrast, are very different: while event-history analysis seeks the causes of phenomena, sequence analysis explores the paths followed without looking for the reasons for the underlying processes that generate the paths (Robette and Bry 2012). Individual characteristics are thus of little value in this approach, apart from the sequence of events and certain characteristics preceding the sequence analyzed. For instance, the 2001 survey by Lelièvre on "Event histories and contact circle," covering a sample of cohorts born between 1930 and 1950, applied a sequence analysis of occupational trajectories of mothers and daughters in order to compare them (Robette et al. 2012). However, there exist some attempts to look at how individual factors may explain the observed heterogeneity between sequences (Studer et al. 2011).

This approach is largely based on non-parametric methods that make no assumptions about the underlying process over a lifetime. Its goal is to explore and describe the course of events as a whole, without worrying about the risk of knowing the events or their determinants. There have also been recent attempts at a Bayesian extension of the sequence approach with the aid of Hidden Markov Models (Bolano 2014; Helske et al. 2018) or variable length Markov models (Gabadinho and Ritschard 2016).

Sequence analysis, in turn, raises a new set of problems that differ from those of event-history analysis. The first problem is mainly a technical one, the others being more theoretical.

The first problem concerns the metric used in social sciences. The OM metric was imported from information theory and molecular biology, where it is fully justified by the basic assumptions. In the social sciences, however, the structure of sequences is far more complex and the metric used is less self-evident. As Wu (2000, p. 46) notes:

> 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, if we interpret substitutions as transitions, assigning the same cost for a substitution from employment to non-employment and for the substitution from non-employment to employment seems scarcely plausible. Bison (2009) clearly shows, using simulations, that different substitution costs yield inconsistent results. As a result, we can find regularities even when none exist (Bison 2014).

To solve these problems, a number of generalizations of the OM method have been proposed, such as variable substitution costs, different distance measures, spell-adjusted measures, non-alignment techniques, and monothetic divisive algorithms. However, with the increase in the number of distances and costs, their comparability becomes increasingly difficult. We do have some comparisons between different metrics, but the only studies comparing a large number of metrics using a set of artificial sequences are those of Robette and Bry (2012) and Studer and Ritschard (2016). In the former study, the authors did not try to find the best metric but "rather to unravel the specific patterns to which each alternative is actually more sensitive" (p. 2). Although they found differences between the results obtained with different metrics, "the main patterns they conceal will be uncovered by most of the metrics" (p. 14). However, these differences exist, and Bison's inconsistent results leave the comparability problem largely unsolved.

The second problem lies in the use of cluster analysis to detect classes of sequences. This classification method was used long before sequence analysis, for example by a psychologist (Tryon 1939) for manual calculation. The advent of computers spurred the development of many methods to detect significant groups, but also raised numerous complications, which we shall summarize here.

One of the most important criteria for a good classification is the number of groups that should exist in a given study. Unfortunately, when the classification criterion is plotted against the number of groups, in most cases, there is no "sharp

step" that we can use to determine the ideal number of classes. The choice becomes highly subjective (Everitt 1979). The assessment of the validity and stability of the clusters found using different approaches is equally problematic. Regrettably, there are few validity tests for these approaches, and even fewer tests of their social significance. A recent comment (Byrne and Uprichard 2012, p. 11) concludes: '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 view the search for causes as obsolete. This raises a third problem: "whether the clusters obtained under this method might be an artifact or something else social" (Wu 2000, p. 51). While unobserved heterogeneity was a significant problem in event-history analysis, here even observed heterogeneity creates difficulties. As sequence analysis tries to capture trajectories as a whole, the only characteristics that can be introduced are the ones measured before the start of the trajectory. Introducing characteristics measured later, or time-dependent characteristics, will raise a host of conceptual problems that are hard to solve. In section three below, however, we shall see that new attempts to combine event-history and sequence analysis may offer a partial solution to these problems.

A fourth problem is linked to the fact that sequence analysis—unlike event-history analysis—cannot fully handle censored observations (Wu 2000, p. 53; Studer et al. 2018). Such a limitation entails the exclusion of incomplete trajectories and confines us to a study of the past. For example, as French retirement age was 65 years at the time of the 3B survey, a sequence analysis of occupational careers would have to be confined to persons born between 1911 and 1916, i.e., one-quarter of the sample.

Sequence analysis is intended to allow a description of trajectories in terms of classes, meant to reflect types of social behavior adopted by groups of individuals. The fifth problem raised is that the meaning of these behaviors is not as clear as one might imagine. First, as an individual is assigned to only one type, the resulting classification is very narrow, whereas we know that an individual may in fact be assigned to a large number of groups such as family, business firms, organizations, and contact circles. These groups are real entities, whereas the classes obtained with sequence analysis are open to question. Consequently, what are the grounds for believing in the existence of these types? Abbott and Tsay (2000, p.27) argue that sequence methods "would find this particular regularity because people in particular friendship networks would turn up in groupings of similar fertility careers". Their argument, however, assumes that data on these individuals' friendship networks are available simultaneously with data on their fertility history. To the best of my knowledge, however, there are no examples showing the congruence of cluster results with friendship networks, but only examples of impact of trajectories on personal network (Aeby et al. 2017).

More recently, several authors have similarly argued that network analysis may be a valuable tool for solving these problems. Bison (2014), for example, suggests converting individual sequences into network graphs. While this method makes it possible "to bring out career patterns that have never previously been observed" (p. 246), it has major limitations. The most important one, advanced by Bison (id),

> that creates the greatest 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 stay in the purely descriptive field of sequence analysis, this characteristic is a genuine hindrance. More recently, Cornwell (2015) goes further and devotes an entire chapter to "Network methods for sequence analysis" (p. 155–209). While some methods used in network analysis may be useful in sequence analysis, it is important to grasp the difference between the goals of the two approaches. The main goal of sequence analysis, as noted earlier, is to understand a life history as a whole and to identify its regularities and structures. Network analysis, as we shall see in the fourth subsection of this section, is focused on understanding the relationships between entities (individuals, or more general levels of "collective agency") and to see how changes at each level drive changes at other levels. We suggest a solution to this problem in the final synthesis of the third section.

2.3 A Level Based Approach

While the two preceding analyses operated at a given aggregation level, we shall now introduce the effects of multiple levels on human behavior. These methods derive from the hierarchical models used in biometrics and population genetics since the late 1950s (Henderson et al. 1959). They were then applied in the social sciences—in sociology by Mason et al. (1983) and in education science in 1986 by Goldstein (2003).

The simplest solution is to incorporate into the same model the individual's characteristics and those of the groups to which (s)he belongs. These "contextual" models differ from cross-sectional models, which explained 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 individual level. It now acts not as a substitute, but as a characteristic of the sub-population that will influence the behavior of one of its members. Simultaneously, we remove the atomistic fallacy, as we take into consideration the context in which the individual lives.

However, contextual models impose 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 individual members of a group behave independently of one another. In practice, the risk incurred by a member of a given group depends on the risks encountered by the group's other members. Overlooking this intra-group dependence biases the estimates of the variances of contextual effects, generating excessively narrow confidence intervals. Moreover, for individuals in different groups, the log-odds cannot vary freely but are subject to tight constraints (Loriaux 1989; Courgeau 2007).

Multilevel models offer a solution to this double problem. By incorporating different aggregation levels into a single model, they generalize the usual regression models. The basic assumption is that the groups' residuals are normally distributed. The analysis can thus focus exclusively on their variances and covariances, but may introduce individual or group characteristics at different levels.

Multilevel analysis no longer focuses on the group, as in the aggregate approach, 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 (Franck 1995, p. 79):

> 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.

We can finally say that this approach regards "a person's behavior as dependent on his or her past history, viewed in its full complexity, but it will be necessary to add that this behavior can also depend on external constraints on the individual, whether he or she is aware of them or not" (Courgeau 2007, p. 79). It can be seen as complementary of event-history analysis but is less linked to sequence analysis.

This approach, however, requires new types of surveys to define and capture the various levels to examine (Courgeau 2007). It has been used in biometrics, population genetics, education science, demography, epidemiology, economics, ecology, and other disciplines. Its methods are basically semi-parametric but can take non-parametric forms, as in factor analysis models.

Although some of these models use the frequentist paradigm, they generally adopt the Bayesian paradigm in order to deal effectively with nested or clustered data (Draper 2008). However, as Greenland (2000) notes, the multilevel approach makes it possible to unify the two paradigms, leading to an empirical Bayes estimate.

But again new problems arise: the three first ones are technical while the last one is mainly structural.

As group characteristics, the multilevel approach often uses mean values, variances or even covariances of group members' characteristics. The first problem is that we must go beyond this approach, for we need a fuller definition of the aims and rules prevailing in a group in order to explain a collective action. What are the mechanisms of social influence that permit the emergence of a collectively owned social capital in different contexts—a capital that "is more than the sum of the various kinds of relationship that we entertain"? (Adler and Kwon 2002, p. 36).

The second problem is that "independence among the individuals derives solely from common group membership." (Wang et al. 2013, p. 125). In fact, the groups are generally more complex. For example a family, generally treated as a simple group, is composed of parents and children, who can 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. As a result, we take into account the interactions between group members and their changes over time in order to fully incorporate their social structure. In the next subsection, we discuss how a multilevel network approach makes it possible to avoid this problem.

The third problem stems from the difficulty of defining valid groups. It leads to the use of geographic or administrative groupings that often have little impact on their inhabitants' behavior. However, by observing existing networks through more detailed surveys, such as those included in the Stanford Large Network Data Set Collection, we should be able to avoid using these unsatisfactory groupings.

The fourth problem is that, while multilevel analysis enables us to incorporate a growing number of levels that constitute a society, it continues to focus on only one of these levels—an event, an individual or a group. As a result, this "approach assumes that links between groups are non-existent." (Wang et al. 2013, p. 1). On the contrary, it is important to take the analysis further by trying to identify the interactions that necessarily exist between the various levels. In Franck's words (1995, p. 79): "the point now is to determine how the different stages or levels connect, from top to bottom and from bottom to top." We shall now see how the analysis of social networks allows us to solve this problem.

2.4 A Network Based Approach

While earlier examples exist, research on social networks effectively began with the work of the sociologists Moreno and Jennings, particularly with a paper (1938) in which they used the term "network theory" and proposed statistics of social configurations. Until the 1970s, however, while research teams in various social sciences worked on network analysis, no cumulative theory resulted (Freeman 2004). Social networks did not begin to be regarded as a full-fledged research field until the 1970s and 1980s. The development of structural models introduced by White et al. (1976) and Freeman (1989) made it possible to examine the interdependent relationships between actors and the similar relationships between actors' positions in the different social networks.

The principle of this approach is to "identify different levels of agency, but also intermediary levels and social forms (such as systems of social niches and systems of heterogeneous dimensions of status), and relational infrastructures that help members in constructing new organizations at higher levels of agency and in managing intertwined dilemmas of collective actions" (Lazega and Snijders 2016, p. 360). It permits to answer to the two last structural problems of sequence analysis, in introducing networks, and the last one of multilevel analysis, in introducing the interactions which exist between the various levels.

This approach rests on robust mathematical foundations. These, however, differ substantially from the previous ones, as the assumption that observations of individuals are independent no longer holds: network analysis argues that units do not act independently but influence each other. The use of graph theory and matrix analysis is important in this field (Wasserman and Faust 1994). Many disciplines—and not only the social sciences—have adopted this approach. They include information science, computer science, management, communication, engineering, economics, psychology, political science, public health, medicine, physics, sociology, geography, and demography.

More recently, we have seen the development of a multilevel network analysis that has provided the link with multilevel analysis. While network theory generally analyzes one given level, the newer approach examines not only the networks that exist at different levels but also the links between levels. It has led to major extensions of existing models of social structures, with networks as dependent variables. One class of models tries to "reveal the interdependencies among the micro-, macro-, and meso-level networks," (Wang et al. 2013, p. 97), the meso-level being defined here as between nodes of two adjacent models". They generalize graph models for multiple networks. A second category of models accommodate "multiple partially exchangeable networks for parameter estimation, as well as pools information for multiple networks to assess treatment and covariate effects" (Sweet et al. 2013, p. 298). Often called hierarchical network models, they are a generalization of the multilevel models described in the previous section. A third type of model "partition[s] the units at all levels into groups by taking all available information into account and determining the ties among these groups." (Žiberna 2014, p. 50). It is a generalization of classical blockmodeling developed for relationships between individuals.

Like the multilevel approach, many of these models use Bayesian estimators—which offer many algorithmic advantages, particularly for non-nested data structures—and Markov Chain Monte Carlo (MCMC) algorithms. They use the frequentist paradigm as well as the epistemic paradigm, producing more general estimators of the empirical Bayes estimator type (Greenland 2000).

This approach requires surveys capable of capturing different levels simultaneously. For example, a survey on relationship networks captured the family, occupational relationships, friendly relationships, and memberships in various organizations and groups for individuals living in a rural area (Courgeau 1972). A network analysis of this survey (Forsé 1981) used a complete diagram of acquaintance networks to construct "sociability" groups distinguished by social and demographic characteristics. Other examples of more restricted networks include biomedical research networks and a secluded monastery (White et al. 1976), as well as larger networks such as those found in the Stanford Large Network Data Set Collection, which comprises social networks, citation networks, collaborative networks, Internet networks, and so on (Leskovec et al. 2009).

What new problems will this approach now encounter? They are now mainly technical ones.

An initial problem is the difficulty of capturing the ties between individuals or in collecting available data on the subject. To begin with, the ties will never be exhaustive, and the many reasons for their limitation complicate their study. Very often, such surveys can capture only a limited number of ties, and the number may vary substantially between surveys. There is also an ambiguity about how to designate ties: the term "best friends" may have a different meaning from "friends most frequently met" or "most trustworthy person." While a survey may ask a respondent for information on different kinds of relationship networks such as family, friends or work colleagues, an existing data collection, such as people linked on Facebook, will not allow this distinction. Some respondents may even report more connections with popular, attractive or powerful persons than they actually maintain.

A second problem is that network clusters are generally created by the researcher rather than pre-existing. The method used to create them requires many decisions that are hard to take in a truly scientific way. As Žiberna (2014, p. 50) noted:

> 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.

While this statement relates more specifically to Žiberna's blockmodeling approach, it also applies to the more general multilevel network approach. In both cases, the researcher must decide whether to include or exclude people from a given network, merge or divide network clusters, and so on. Such decisions are needed to allow statistical analysis later on.

A third problem is the difficulty of introducing individual or network characteristics in the study. This can be done only by using the hierarchical network model. But, even in this case, few data sets give measures of the effects of characteristics or measures of network structure (Sweet et al. 2013). These characteristics may be individual, network-specific, tie-specific, or a combination of the three.

A fourth problem concerns the introduction of time in these studies. Here as well, very few surveys enable us to observe changes in networks over time. Some multi-wave surveys capture network structure at different times. Lazega et al. (2011) used a three-wave survey to show that an organization's structure remains the same regardless of its membership's turnover. However, we need more detailed surveys on the changes in networks over a long, continuous period in order to study the changes that may occur, up to and including the end of a network.

We can conclude this examination of the problems and challenges of multilevel network analysis with the following quotation (Lazega and Snijders 2016, p. 260):

> 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.

Arguably, these problems should be seen more as a challenge for future research than as insuperable difficulties.

3 Toward a Synthesis

After describing and assessing four approaches—with different goals—to understanding human behavior, let us now see if we can give a more synthetic view of them. We begin by examining two basic concepts without which no social science would be possible.

The first concept is the creation of an abstract fictitious individual, whom we can call a statistical individual as distinct from an observed individual. While for Aristotle (around 350 BC, Book I, Part 2, 1356b) "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 by setting aside the observed individual—too complex for study—and using statistics on a small number of characteristics, yielding a statistical individual. In Courgeau's words (2012, p. 197):

> 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 statistical individual having been thus defined, the key assumption that allows the use of probability theory here is that of exchangeability (de Finetti 1937), which we can formulate simply as follows: n trials will be said to be exchangeable if the joint probability distribution is invariant for all permutations of the n units. Social scientists routinely use exchangeability for the residuals obtained, taking into account the various characteristics included in their analysis. In so doing, they distinguish the statistical individual from the observed individual.

The second concept is the statistical network, as distinct from observed networks. It was introduced more recently, by Coleman (1958). While observed networks may be as diverse as the different kinds of ties existing between individuals—consistently with Aristotle's comment on individuals—statistical networks are obtained from an analysis of ties between individuals along with and the choice of criteria to circumscribe the ties. Here as well, the basic assumption that allows the use of probability theory is that of the exchangeability of networks and the individuals that compose them, taking into account the characteristics introduced at each level.

It is interesting to compare these two concepts with the contexts proposed by Billari (2015) to explain population change, namely, the micro- and macro-level contexts. In fact, Billari clearly recognizes the abstract concept of statistical individual—the same concept proposed here—as the basis of the micro-level context. For the macro-level context, however, he only proposes to examine how "population patterns re-emerge from action and interaction of individuals" (p. S13), without fully recognizing the abstract concept underlying the interactions: the statistical network, which makes it possible to flesh out this macro-analysis. As an example, we have already seen how multilevel analysis reconciles the macro- and micro-level results.

Once these two main concepts are defined, we can see that the study of time between events and the study of event sequences are directly connected to the same concept of statistical individual. Despite the earlier-noted difference in their approaches to this individual, we can regard them as two complementary ways to study the individual. Furthermore, some recent studies combine the advantages of the two approaches by modeling "interaction between macro-institutional configurations and individual life-course trajectories" (Studer et al. 2018). The definition of event-history analysis, already given on the first subsection of the second section of this chapter easily extends to sequence analysis. The itinerary is followed event after event in the first approach and with more complex sequences of events in the second approach.

Similarly, we can see that the contextual, multilevel, and multilevel network approaches are simultaneously connected to the same concept of statistical network. They also seem complementary. We can say that contextual and multilevel analysis focuses on attributes of both individuals and levels, whereas network multilevel analysis focuses on relationships combining the different levels. The paradigm offered for the contextual and multilevel approach is "to explain dependent variables by models containing multiple sources of random variation and including explanatory variables defined as aggregate or other higher-order units" (Lazega and Snijders 2016, p. 3). They can easily extend this paradigm to the network-based approach, with the additional specification that it "means analyzing separately, then jointly, several models of collective agency" (p. 4).

Interestingly, multilevel approaches may be seen as complementing event-history analysis by introducing the effects of membership of different levels on individual behavior. Similarly, multilevel network analysis may be seen as complementary to sequence analysis. This proximity may explain why Cornwell (2015) tries to introduce network analysis methods in sequence analysis. However, sequence methods rely mostly on a grouping of statistical individuals determined by personal criteria, while network methods introduce statistical networks from the outset.

The problems encountered when using one of the four approaches above are easily solved by simultaneously examining the statistical individual and the statistical network by means of a more general biographical multilevel network analysis. As noted earlier, such an approach avoids the risk of atomistic or ecological fallacy by using a synthesis of holism and methodological individualism. It also avoids having to choose between Bayesian and frequentist probability through the use of a more general compromise on confidence distributions (Schweder and Hjort 2016), paving the way for a more satisfactory statistical inference. By introducing networks that yield a better understanding of human behavior, it offers solutions to several problems posed by unobserved heterogeneity. It is also likely that a number of problems involved in sequence analysis—such as the choice of metric, cluster analysis, and the question of whether the groups formed actually exist—can be solved by undertaking more complex surveys on social networks. These would enable us to replace theoretical clusters with real networks of individuals linked together by existing social forces. Similarly, the main problems raised by multilevel analysis could be solved more easily by multilevel network analysis, such as the use of a Multilevel Social Influence (MSI) model (Agneessens and Koskinen 2016) to explain the emergence of social capital, and the use of Exponential Random Graph Models (ERGMs) to show that within-level network structures depend on network structures at other levels (Wang et al. 2016).

Lastly, we believe that the problems posed more recently by multilevel network analysis should be seen as a challenge for future research rather than as insuperable difficulties. For example, such an analysis will reach its full potential when truly longitudinal observations of the multiple levels analyzed become available, providing a combination of individual and network event histories. Collecting data and providing valid statistical approaches to solve this problem will be necessary and appear to be a challenge for such a biographical multilevel network analysis.

4 Conclusion

If we define a scientific approach solely by its methods, we inevitably adopt a partial view of the core of the approach. We must now 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 can be traced back to Bacon in 1620 (Bacon et al. 2000, XIX, p. 36):

> There are, and can be, only two ways to investigate and discover truth. The one leaps from senses and particulars to the most general axioms, and from these principles and their settled truth, determines and discovers intermediate axioms; this is the current way. The other elicits axioms from sense and particulars, rising in a gradual and unbroken ascent to arrive at last at the most general axioms; this is the true way, but it has not been tried.

Bacon calls the second approach induction, not in the meaning later given to the term by Hume and his empiricist tradition—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, biological or social.

It is important for the social sciences to begin by observing and measuring social facts, for this measurement, far from being of secondary importance, makes it possible to assess the "potentialities" of a social fact (Courgeau 2013). Next, instead of relying on often arbitrary hypotheses, 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. 2017).

While we can argue that individuals each have an unlimited and unknowable number of characteristics with their own freedom of choice, social science can show that they are born in a given society with its rules and laws, which restrain their freedom, and that they are subject to biological laws, which are the same for all humans. This is what allows the existence of a social science that takes into account a limited number of characters and is based on a set of concepts without which these characters would be inconceivable or impossible (Franck 2002).

We should like to conclude by emphasizing the following point. We have more often viewed the social sciences 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 believe it is possible to construct a new formal object that can explain certain properties of human societies—an object that transcends existing disciplines and allows their synthesis.

Acknowledgements I thank the three anonymous referees for their detailed and thoughtful comments on an earlier draft of the manuscript, and Jonathan Mandelbaum for its English translation.

References

Aalen, O. (1975). *Statistical inference for a family of counting processes*. Ph.D. thesis, Institute of Mathematical Statistics, Copenhagen.

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: A Journal of Quantitative and Interdisciplinary History, 16*(4), 129–147.

Abbott, A. (1984). Event sequence and event duration: Colligation and measurement. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 17*(4), 192–204.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21*(1), 93–113.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology. *Sociological Methods & Research, 29*(1), 3–33.

Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. *The Academy of Management Review, 27*(1), 17–40.

Aeby, G., Gauthier, J.-A., Gouveia, R., Ramos, V., Wall, K., & Cesnuityte, V. (2017). The impact of coresidence trajectories on personal networks during transition to adulthood: A comparative perspective. In V. Cesnuityte, D. Lück, & E. Widmer (Eds.), *Family continuity and change: Contemporary European perspectives* (pp. 211–242). London: Palgrave Macmillan.

Agneessens, 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 E. Lazega & T. A. Snijders (Eds.), *Multilevel network analysis for the social sciences: Theory, methods and applications*, (pp. 81–105). Cham: Springer.

Aristotle (350 BC). *Rhetoric*. Translated by W. Rhys Roberts (http://classics.mit.edu/Aristotle/rhetoric.html).

Bacon, F., Jardine, L., & Silverthorne, M. (2000). *The new organon* (Cambridge texts in the history of philosophy). Cambridge: Cambridge University Press.

Billari, F. C. (2015). Integrating macro- and micro-level approaches in the explanation of population change. *Population Studies, 69*(sup1), S11–S20.

Bison, I. (2009). OM matters: The interaction effects between indel and substitution costs. *Methodological Innovations Online, 4*(2), 53–67.

Bison, I. (2014). Sequence as network: An attempt to apply network analysis to sequence analysis. In P. Blanchard, F. Bülmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 231–248). Heidelberg: Springer.

Bolano, D. (2014). Hidden Markov models: An approach to sequence analysis in population studies. In *Annual Meeting of the Population Association of America, Boston*, 1–3 May 2014.

Bretagnolle, & Huber-Carol (1988). Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics, 15*(2), 125–138.

Byrne, D., & Uprichard, E. (2012). Introduction. In D. Byrne & E. Uprichard (Eds.), *Cluster analysis* (Vol. 2, pp. vii–xii). London: Sage Publication Ltd.

Coleman, J. (1958). Relational analysis: The study of social organizations with survey methods. *Human Organization, 17*(4), 28–36.

Cornwell, B. (2015). *Social sequence analysis: Methods and applications*. Cambridge: Cambridge University Press.

Courgeau, D. (1972). Les réseaux de relations entre personnes. etude d'un milieu rural. *Population, 27*(4–5), 641–683.

Courgeau, D. (1999). L'enquête "triple biographie: Familiale, professionnelle et migratoire". In G. de réflexion sur l'approche biographique (Ed.), *Biographies d'enquêtes* (pp. 59–74). Paris: INED.

Courgeau, D. (2007). *Multilevel synthesis: From the group to the individual*. Dordrecht: Springer.

Courgeau, D. (2012). *Probability and social science: Methodological relationships between the two approaches*. Dordrecht: Springer.
Courgeau, D. (2013). La mesure dans les sciences de la population. *Cahiers philosophiques, 135*(4), 51–74.
Courgeau, D., & Lelièvre, E. (1992). *Event history analysis in demography*. Oxford: Clarendon Press.
Courgeau, D., & Lelièvre, E. (1997). Changing paradigm in demography. *Population, 9*, 1–10.
Courgeau, D., Bijak, J., Franck, R., & Silverman, E. (2017). Model-based demography: Towards a research agenda. In A. Grow & J. Van Bavel (Eds.), *Agent-based modelling in population studies: Concepts, methods, and applications* (pp. 29–51). Cham: Springer.
Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187–220.
Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. London: Chapman and Hall.
de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Pointcarré, 7*(1), 1–68.
Doob, J. L. (1953). *Stochastic processes*. New York: Wiley.
Draper, D. (2008). Bayesian multilevel analysis and MCMC. In J. Deleeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 77–140). New York: Springer.
Durkheim, É. (1895). *Les règles de la méthode sociologique*. Paris: Alcan.
Everitt, B. S. (1979). Unresolved problems in cluster analysis. *Biometrics, 35*(1), 169–181.
Forsé, M. (1981). Les réseaux de sociabilité dans un village. *Population, 36*(6), 1141–1162.
Franck, R. (1995). Mosaïques, machines, organismes et sociétés. examen métadisciplinaire du réductionnisme. *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, 31*(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: Kluwer Academic.
Freeman, L. C. (1989). Social networks and the structure of experiment. In L. C. Freeman, D. R. White, & A. K. Romney (Eds.), *Research methods in social network analysis* (pp. 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, BC: BookSurge.
Gabadinho, A., & Ritschard, G. (2016). Analysing state sequences with probabilistic suffix trees: The PST R library. *Journal of Statistical Software, 72*(3), 1–39.
Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.
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, 29*(1), 158–167.
Helske, S., Helske, J., & Eerola, M. (2018). Combining sequence analysis and hidden Markov models in the analysis of complex life sequence data. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications* (Life course research and social policies). Berlin: Springer (this volume).
Henderson, C. R., Kempthorne, O., Searle, S. R., & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics, 15*(2), 192–218.
Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer.
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: Wiley.
Lazega, E., 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*(1), 129–144.
Lazega, E., & Snijders, T. A. B. (Eds.) (2016). *Multilevel network analysis for the social sciences: Theory, methods and applications*. Heidelberg: Springer.

Leskovec, J., Lang, K. J., Dasgupta, A., & Mahoney, M. W. (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.

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, Vol. 1987, pp. 333–368). Louvain-la-Neuve: Ciaco.

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual analysis through the multilevel linear model. *Sociological Methodology, 14*, 72–103.

Moreno, J. L., & Jennings, H. H. (1938). Statistics of social configurations. *Sociometry, 1*(3/4), 342–374.

Robette, N., & Bry, X. (2012). Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 116*(1), 5–24.

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* (pp. 395–418). INED.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357.

Schweder, T., & Hjort, N. L. (2016). *Confidence, likelihood, probability: Statistical inference with confidence distributions*. Cambridge: Cambridge University Press.

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society, Series A, 179*(2), 481–511.

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research, 40*(3), 471–510.

Studer, M., Struffolino, E., & Fasang, A. E. (2018). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. *Sociological Methodology* (First Published Online).

Sweet, T. M., Thomas, A. C., & Junker, B. W. (2013). Hierarchical network models for education research. *Journal of Educational and Behavioral Statistics, 38*(3), 295–318.

Trussell, J. (1992). Introduction. In J. Trussell & R. Hankinson (Eds.), *International Studies in Demography* (pp. 1–7). Oxford: Clarendon Press.

Trussell, J., & Richards, T. (1985). Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure. In N. B. Tuma (Ed.), *Social and behavioral science series*(Vol. 15, pp. 242–276). San Francisco, CA: Jossey-Bass.

Tryon, R. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Ann Arbor: Edwards brother.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16*(3), 439–454.

Ville, J. (1939). *Etude critique de la notion de collectif*. Paris: Gauthier-Villars.

Wang, P., Robins, G., & Matous, P. (2016). Multilevel network analysis using ERGM and its extension. In E. Lazega & T. A. Snijders (Eds.), *Multilevel network analysis for the social sciences* (pp. 125–143). Cham: Springer.

Wang, P., Robins, G., Pattison, P., & Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks, 35*(1), 96–115.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

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.
Wu, L. L. (2000). Some comments on "Sequence analysis and optimal matching methods in sociology: Review and prospect". *Sociological Methods & Research, 29*(1), 41–64.
Žiberna, A. (2014). Blockmodeling of multilevel networks. *Social Networks, 39*, 46–61.