STATISTICS IN TRANSITION-new series, Summer 2012
Vol. 13, No. 2, pp. 365—386

STATISTICS AND SOCIOLOGY: THE MUTUALLY-SUPPORTIVE DEVELOPMENT FROM THE PERSPECTIVE OF INTERDISCIPLINARIZATION OF SOCIAL RESEARCH[1]

Włodzimierz Okrasa[2]

ABSTRACT

Statistics emerged as a scientific discipline and has been developed as such, especially extensively over the past century, not only due to an extraordinary service it provides to other disciplines but also thanks to ideas, questions and approaches originally formulated in different fields of empirical research, including sociology, which also contributed to statistics. The confluence of developments in these two disciplines (statistics and sociology) seems to be one of the most successful and beneficial for both of them. Yet, it has become a focus of systematic reflection only recently. The aim of this paper is to make a concise overview of the logical scheme of this interaction and to stress the importance of the counterfactual causal modeling being currently under constant refinement. A more explicit formula of interdisciplinarization that underlies such an interaction anyway would add to overcoming the methodological challenges it poses to either discipline. While contributing to the advancement of 'cause-and-effect' oriented quantitative sociology this would enhance methodology of social science research in general.

I. The key aspects of the interplay between statistics and sociology

In most of historical essays on the development of statistics that account for other disciplines' contributions to it – for instance, such as of S. Stigler's (1986) narrations on how statistics arose from the interplay of mathematical concepts with astronomy, geodesy, experimental psychology, genetics, and sociology – the last one, sociology, is typically being mentioned among the leading contributors. On the one hand, studying impact of statistics on other scientific fields, might allow us to look at the history of science through the window of its own progress

---

[1] This article is based on a paper presented at the Congress of Polish Statistics: 100th Anniversary of the Polish Statistical Association, Poznan, April 18-20, 2012.

[2] Central Statistical Office (GUS) and the University of Cardinal Stefan Wyszynski in Warsaw (UKSW).

(as suggested by Fienberg, 1992). And such a perspective seems to be an obvious way to trace the role statistics plays in the growing process of interdisciplinarization of research. On the other hand, adopting a more disciplinary-oriented view – as in this paper, with its primary interest in the question of whether and how has sociology played a part in the development of modern statistics – the issue goes beyond historical aspects. This paper focuses on the elements of confluence or 'commonalities' in the methodological advancement of both disciplines, statistics and sociology, and how this interplay contributes to the interdisciplinarization of social research.

Interdisciplinarization of empirical research seems to be an intrinsic effect of employing statistics. As J. Neyman used to say, "[S]tatistics is the servant to all sciences" (cf. Chiang), though in rather different way for each discipline. The nature of this service is, to a large extent, shaped by methodological framework and an overall research paradigm in particular field of science. Such a view can be justified on the historical ground – in science *en globe*, and also within the picture narrowed to a social science field of research. It allows for identifying a discipline-specific subject matter, on the one side, and statistical issues on the other, just providing a needed base for interaction between them, especially between sociology and statistics. Several statistical methods (developed within particular discipline) have been fruitfully combined on the ground of sociological research. With some time lag, however – for instance, *path analysis* has been imported from biometricians, *factor analysis* from psychometricians, and *structural modeling* from econometricians.

The process of mutual influence between statistics and sociology is taking place across different planes, with diverse intensity and character over the time. A typology of respective approaches can be based on specification of the main object of the focus - being it either (problem-oriented) statistical research method or type of data-oriented approach, or an approach focused on the type of analysis, or their cross-roads between disciplines. The development in quantitative methods in sociology will be discussed together with improvements in statistical data production during the post-war evolution of both disciplines. Looking back, and employing the four phase perspective of historical development of statistics since the mid of XVII to the mid of XX centuries, as it is typically proposed (cf., Fienberg, op cit.), it should be noted that sociology has entered into research interaction with statistics only during the third phase that spanned from 1820 to 1900. Owing it mainly to A. Quetelet, who is considered the father of quantitative social science (e.g., Fienberg, op. cit.) and this phase is called 'the socialization of statistics' due to his concept of "the average man" and work he did on fitting of the normal distribution (that was originally used by Gauss and Laplace in astronomy) to several social data sets. For instance, to the distribution of body heights measured in population of conscripts, and other data produced by censuses and within 'social physics' (Fienberg, op cit., p. 216), as the new field of research was called before the term 'sociology' was coined by A. Comte.

According to its founder, sociology was grounded in statistical data, what has soon been demonstrated in a spectacular way by Durkheim's studies of suicides.

In his more disciplinary-oriented narration on quantification in sociology, Paul Lazarsfeld (1961) talked about two-way path ('the two roots' in his words – op. cit., p. 283) that subsequently led the newly emerged discipline to its so much data-loaded version: English 'political arithmeticians' (W. Petty in 17th century) and German universities engaged in description of state's features. Originally it was made in non-quantitative way, as it did G. Achenwall, who created the term 'statistics' in 18th century; however, according to Lazarsfeld, it was Herman Conring who already in 17th century elaborated a system of tables (rows and columns), laying ground for 'university statistics' (*ibidem*, p. 286-291). Within each of these two paths efforts have been made to generate statistical knowledge that was indispensable for what Foucault called 'correct government'. Nb. as a university discipline, quantitative sociology was established at the end of 19th century, with the appointment of Franklin H. Giddings as professor of this field at Columbia University (1894)[1]

- **Two-way interaction**: ***statistics in sociology*** and ***sociology of statistics.*** Statistics, especially official statistics – meant as figures such as consumer prices, the unemployment rate, economic indicators, or data on crime, divorce, and poverty, being produced by government – affect both public and individual decisions in many ways. In doing so, official statistics do not only holds a mirror to reality but - as it was noted by Alonso and Starr (editors of the remarkable collection of papers entitled "The Politics of Numbers" (1988) - they also reflect presuppositions and theories about the nature of society. Such a reciprocal influence might be described with Winston Churchill's metaphor that *we shape our buildings and then they shape us* (Alonso and Starr – op cit., p. 3). And it does involve the sociological question of how quantification and statistical algorithms contribute to shaping the social world, as asked by Desrosierès (2011), who proposed a framework for analyzing this issue based on the idea of two-dimensional construction. The first concentrates on (i) how society and the economy are being conceptualized; (ii) modes of public action, and (iii) different forms of statistics and of their treatment. The second encompasses the five models of state which differ in terms of specific to each of them statistical instruments, such as (Desrosierès, op cit. p 42-45): (a) the French-style *engineering state* and centrally-planned economies – statistics specific to it focuses on production in physical quantity, input-output table, material balance and demography; (b) the *classical liberal state,* concerned about equal access to information for all stakeholders – statistics promoting market transparency,

[1] In addition to F. H. Giddings as the first who institutionally brought statistical methods to sociology, the authors identify such pioneers in other disciplines as well: McKeen Cattell in psychology, F. Boas in anthropology, and H. L. Moore in economics (all faculty members at Columbia University, which "provided a conducive setting for the interdisciplinary process of the incorporation of statistical methods into the social sciences" (Camic, Xie, 1994, p. 773.).

and market shares; (c) the *welfare state* concerned with employment and family protection schemes – interest in labor statistics, consumer price indexes, and sampling surveys of households; (d) crisis-related model of the *Keynesian state*, with policies dealing with aggregates (models like of Lawrence Klein's/LINK) – main interest in national accounting, in the figures on consumption and the price index; (e) *the neoliberal state* (that emerged in 1990's) focused on microeconomic dynamics – interest in incentive systems (according to the rational expectations theory), and in statistical indicators to evaluate performance.

The relationship between statistics and sociology that is of interest here is two-fold – methodological and institutional. From the methodological angle, statistics is meant as the means of data generating and analysis; its particularly extensive use observed in such disciplines as psychometrics, econometrics (the term 'sociography' was also proposed by F. Tönnis around the turn-of-the-century, though unsuccessfully). From the institutional standpoint, statistics is a domain of public activity ('public statistics'). One of the means to institutionalize the statistics was census of population, which is traced to ancient societies[1]. Institutionalization is being manifested in the best way by government statistics or official statistics, which itself becomes a research object within *sociology of official statistics*. It aims to shed light on "why and what is being counted – or sometimes even more interestingly – not counted" (as phrased by G. Sternlieb (1973) to whom creation of sociology of statistics is being attributed – see Starr (1987)). A classic framework of sociology of statistics encompasses, according to Starr (op cit.) the five main areas of interest: *system origins* and development – s*ocial organization* of statistical systems, including relationships between the agents involved in the distribution and use of the statistical information – c*ognitive organizations* of statistical system meant as the intellectual construction of presuppositions, rules, categories of classification and methods of measurement used to produce the information by the statistical institutions – *system uses* (how production and distribution of statistical information shape policy-making and society) – and contemporary *system changes*.

- **Approaches 'to historicize' the interplay of statistics and sociology over time.** The above mentioned three (post-war) generations of quantitative methods in sociology which have emerged in connection with the kinds of data they address, can be conceptualized after Raftery as follows: (i) the late 1940s to early 1960s – focus on cross-tabulations and on measures of association and *log-linear models* (the area of statistics to which sociology has contributed the most, following Pearson's and Yule's earlier ideas of cross-product and odds ratio); (ii) the second generation, which began in the

[1] Census was meant in ancient Rome as "a register of adult male citizens and their property for purposes of taxation, the distribution of military obligation, and the determination of political status." – see Wilcox W.F., 1930. "Census", *Encyclopedia of the Social Sciences*, New York: Macmillan.

1960s and continued to early 1980s, dealt with unit-level survey data (facilitated by the availability of high-speed computers), with special focus on *path analysis* and LISREL-type causal models, and *event-history analysis;* (iii) the third generation started in the late 1980s, with methods dealing with 'newer data forms' that comprise data that do not fall into none of the above categories or forms, such as *spatial* or social *network* data, and *texts* or *narratives*. Somewhat more detailed specification of phases in post-war era of statistics-and-sociology methodological interaction, albeit for only until 1980s, was proposed by Bernert (1983): post-war to 1950, after 1950 to late 1960s, and 1980s. [Omitted here is the pre-war era (categorized by Bernert into three periods: before WW1, 1920-30, and 1930s) during which so-called '-tricians' have contributed to statistical methods -biome*tricians* (path analysis), psychome*tricians* (factor analysis), and econome*tricians* (modeling theory) – which next attracted quantitative sociologists; especially those engaged in causal analysis, during the whole post-war period (until 1980s).]

Out of all the data-and-method types of advancements in methodology of social research during the post-war era it perhaps was *sampling survey* that has been of particular interest to sociologists – as 'a telescope on society' (House et al., 2004) – and influenced the interaction between statistics and sociology in a special way. Modern representative sampling method was, according to Kish (1995)[1] and others (e.g., Heeringa et al., 2010) possible only thanks to earlier works of J. Neyman (especially, his path breaking article (1933) that laid the ground for theory of inference along with works of R. Fisher (1935) and his collaborators involved in survey sampling (Cochran, Mahalanobis, Yates, Snedecor). In particular, sociologists were keen to the 'design-based' theory of inference employed to data generated under probability sample (following Neyman's original framework), as alternative to model-based method of inference from data collected under model-dependent sample. [A typology proposed by Hansen et al., (1983) that results from cross-classification of 'sampling plan' (probability and model-based sample) and 'method of inference' (design-based and model-based inference) does not involve non-probabilistic methods of sample selection – quota sampling, convenience sampling, snowball sampling and 'peer-nomination' – which, however, have also attracted the survey practicioners, including sociologists interested in variety of purposive sampling.]

Some of the major issues in quantitative methodology which have marked the post-war era are briefly discussed below in line with the above quoted periodization (i.e., following Raftery – op. cit., taking also into account some suggestions from Bernert, 1983 and Kish, 1995). In addition to the already stated question of mutually-supportive methodological developments in statistics and

[1] According to Kish, an official birth date for survey sampling marks the publication of paper by A. N. Kiaer (1897) on 'representative method' in Bulletin of International *Statistical Institute* (Kish. 1995).

sociology, some remarks will also be added on one of the hottest topic discussed in the literature, namely, the epistemological status of statistical associations (correlation) and causality, with special emphasis on *counterfactual framework* of causation in social research.

II. Three generations of interactional development

1. Association model methodology – *models for cross-tabulations*

Two types of analytical tools, *loglinear models* and *latent class models*, have emerged during the two post-war decades and contributed significantly to better use of categorical data in social research, and subsequently in other disciplines as well. The first one started with data on social mobility; the other with multivariate discrete data (such as indicators of directly unobserved constructs).

- *Loglinear models*: Thanks to them the social mobility tables have been given an appropriate specification in terms of associations – with separation of structural mobility from circular or exchange mobility – in the model originally defined by Birch (1963) as *loglinear model* for the observed counts $\{x_{ij}\}$:

$$\log(E[x_{ij}]) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

where *i*, *j* are index rows and columns, respectively; *u1(i)* and *u2(j)* are the *main effects* for the rows and columns, *u12(ij)* is the *interaction* term.

Since the model's applicability was limited given a large number of parameters required, Duncan (1979) and Goodman (1979) provided a more general approach to modeling the interaction terms using association model. Thanks to describing associations in terms of local odds ratios, Raftery and his colleagues (see Raftery, op. cit., p. 7)) managed to analyze the tables comprising 7,000 cells and five dimensions: (1) father's occupation, (2) offspring's occupation, (3) gender, (4) race, and (5) period. One important result of employing such models (by different authors) was obtaining a better picture of social mobility in westerns societies (e.g., that the mobility has been increasing in USA; that social mobility increased by 1 percent a year in industrialized countries, through the 2nd half of the 20th century). It also allowed clarifying some of social policy relevant issues like relationship between household vulnerability and recurrent poverty (Okrasa, 1999). In order to determine how vulnerability affects the household welfare status over time, the association between the two variables was estimated using a version of the above model for poverty pattern over time (welfare-path or trajectory) and household vulnerability (Okrasa, *ibidem*):

$$\ln(m_{ij}) = \mu + \lambda_i^{\text{welfare-path}} + \lambda_j^{\text{vulnerability}} + BU_iV_j$$

where μ represent the constant and lambdas are the main effects parameters, and *B* is a regression coefficient multiplied by the scores U and V assigned to the cell at i row and j column. As expected, chronically vulnerable households were facing substantially higher risk of becoming long-term poor than non-vulnerable households. The odds of being in poverty for at least one year, rather than remaining outside the poverty during the whole four year period under study were nearly twice as high (1.8) for vulnerable households than for non-vulnerable households. The same odds ratio holds between any two adjacent categories of the poverty pattern over time.

It is worth mentioning that association models – as shown by Goodman (1985) – are related to *canonical correlations* and to *correspondence analysis*. According to Raftery (op.cit.), the use of association models has diffused from sociology to other disciplines, such as epidemiology.

- *Latent Class Models (LCM):* The latent class models were introduced to account for observed associations in multivariate discrete data (Lazarsfeld (1950), Goodman (op. cit.)). To some extent, it was done in analogy to the idea of *factor analysis* for multivariate continuous data, or to *cluster analysis* for the case of binary data (but knowing a priori how many clusters there might be, and that that allocation of an object to a cluster is probabilistic – Bartholomew et al., 2002). In brief (for details see for instance Bartholomew et al., *ibidem*, pp. 235-245) it is assumed that every object (out of n objects consisting a random sample from some population) belongs to just one of the J latent class.

Let $\pi_{ij} = \Pr(x_i = 1|j)$ be the probability that an object from class *j* will answer positively to item *i* ($i = 1, \dots ,p$; $j = 1, \dots , J$); and η_j be the proportion of the population in latent class j (or equivalently the probability that a random selected object from the population belongs to latent class j, for ($j = 1, \dots , J$). The probability of giving a positive response to a particular item is the same for all objects in the same class; given the latent class to which an object belongs, its response to different items is conditionally independent. And η_j be the proportion of the population in latent class j, or equivalently the probability (the *prior* probability) that a random selected object from the population belongs to latent class j, for ($j = 1, \dots , J$). The model is fitted iteratively to obtain maximum likelihood estimates, $\hat{\pi}_{ij}$ of π_{ij} , and $\hat{\eta}_j$ of η_j. Allocation to classes takes place according to (the *posterior*) probability that an object with a particular response pattern falls into a particular class, is: Pr(object is in class $_j| x_1, \dots, x_p$), ($j = 1, \dots ,J$).

It is difficult sometimes to distinguish empirically between a latent class model and a latent *trait* model. For instance, Bartholomew and others indicate on the number of determination parameters – if it is large, a latent class model seems to be more appropriate (*ibidem,* p. 247). Several different versions of the extension of the basic latent class model have been proposed in the literature since its origination until recently – including introduction of the Markov Chain Monte Carlo (MCMC) methods for parameter estimation (one of the features facilitated by substantial growth in power of computers – cf. Bartholomew et al., 2011).

2. Unit-level survey data

The second generation of interaction between statistics and sociology is (according to Raftery, op. cit.) by researchers' efforts to shift from *functional* to *causal* type of analysis.

- *Path analysis* & s*tructural equation models.* Macro-sociologists preoccupied with analysis of stratification process in terms of inter-generational changes in occupational status have moved from linear regression toward a causal interpretation based on Blau's and Duncan's (1967) *path model* as the archetype model of occupational attainment (following an earlier idea of Wright (1921)). It started with the Duncan Socioeconomic Index (SIE) (according to which the prestige score of 45 occupations presented in the 1960 Census, based on the proportion of occupational incumbents who have completed high school and those who have earned more than $10 000, was predicted with high $R^2 = 0.91$ - see Raftery, p. 11). It was extended to a more general methodology of *structural equation models.* According to the Balu and Duncan famous model, which was replicated in various modifications and in different contexts in the literature for decades, two pairs of sequential variables were responsible for variability in occupational status. Namely, respondent's occupation showed to be of an effect of father's education and father's occupation (affecting mutually each other) which subsequently influenced both respondent's education and respondent's first job.

  Decomposition of a total effect into direct and indirect effects in structural model, along with causal interpretation given earlier to them by Blalock (1961) and, on the other hand, with presentation of the measurement model for unobservable latent variables (such as ability or attitude or motivation, etc.) while treating it as an integral part of the 'causal relation' model, led to linear structural relationship methodology (including LISREL-software by Jöreskog (1973)). Its different versions for various types of data – including categorical data, longitudinal data, and multilevel data, sometimes in combination with other models (e.g. graphical models, for instance by computer scientists) – continue to develop as one of the powerful tool for modeling causal type associations in sociological generalizations or theories. However, the causal interpretation of structural models' parameters has being vigorously criticized during the past few decades both by statisticians (Freedman, 1987, Sobel, 1995) and social science methodologists (Morgan and Winship, 2007). But this issue, and the state of the current discussion on modeling causation based on non-experimental design or observed data, represent a separate kind of problem that deserves discussion, and can only be briefly mentioned below.

In order to complete an overview of the methodological innovations which are worth mentioning as exemplifying another aspect of cooperation between statisticians and sociologists (during the second generations of their developments in Raftery's periodization) two other data-suited methods deserve to be brought up: *event history* analysis and *limited dependent variable*s.

- *Event history* methodology for analyzing data on such occurrences as marriages, divorces, births, job changes, going on or off welfare, along with factors influencing such events – e.g., as dropping out school, or falling in (or moving out of) poverty (Okrasa, 1999) – was introduced by Cox (1972). The Cox proportional hazard model has brought together essential elements of two approaches being previously in use for working with event-history data: life table analysis to count for time (from demography) and regression analysis to also count for time-related influencing factors (but troubled with the problem of censoring). Of particular interest to sociologists showed to be a class of discrete-time event-history models (Allison, 1982), extended to such complex areas of study as innovations and social influence due to, among others, using accelerated failure-time models rather than proportional hazards models (Yamaguchi (1994) – cf. Raftery (op. cit., p. 15).
- *Limited dependent variables*, as a method for analyzing categorical dependent variables – a special case of which is binary dependent variable, though nominal and ordinal variables are included too – is another example of a procedure in the development and implementation of which sociologists have been effective since its commencing (following its earlier application in health research) making significant contribution (cf. Long, 1997). *Logistic regression* showed not only to suit better many sociological data – not necessarily with binary dependent variables (e.g., Agresti, 1990) – than does linear regression model. It also is preferred by sociologists over another alternative to linear regression, such as *probit analysis*, due to convenient interpretation of the logistic regression coefficients in terms of odds ratios. From among different extensions of the limited dependent variables, two types of models are perhaps most significant and useful. One was developed by econometricians in the context of evaluation research (such as sample selection model with two-stage estimator – see Heckman, 1979). Another type of model was elaborated in the context of compositional data such as analysis of household expenditures represented in form of vector of shares of particular consumer items (summing up to one).

3. Causality and spatial data

According to the adopted here perspective (after Raftery, op. cit.) on the confluence of statistical and sociological methodologies based on types of data emerging as specific to a given period of time, in the last (third) phase of the post-war era the following new kind of data showed to be of special interest to sociologists: social networks and spatial data – textual and qualitative data –

narrative and sequence analysis. For each of them new statistical methods have been proposed, but given the limitation of space it is not possible to characterize them here. Therefore, leaving aside all other types of data some remarks confine to spatial data and spatial analysis.

Spatial methodologies. Spatial data and analysis are, surprisingly enough, considered to be less advanced in sociology than in other disciplines (e.g. Urry). This contrasts with the fact that not only social data are inherently spatial – since all social phenomena take place in space – but also because they seem to provide a natural basis for integration of social science (Goodchild et al., 2000) and for interdisciplinarization of social research (Okrasa, 2012). According to Fischer and Getis (2010, p.2) there are disciplinary distinctions in the interest in particular types of spatial data: sociologists share strong interest with anthropologists and geographers in point and area (chloropleth) data; economists and political scientists are especially fond of time series data; environmentalists use remotely sensed spatial data; planners, especially in domain of transport, prefer to work with network data. These distinctions are, to some extent, reflected in different 'schools of thought' on spatial analysis methodology. There are four such schools identified (*ibidem*, pp. 3-6), which however overlap with each other in terms of certain concepts and statistics (such as spatial autocorrelation statistics), as follows: (i) *exploratory spatial data analysis (ESDA*) which is an extension of so-called Tukey-type data exploration (or Tukey's exploratory data analysis /EDA) to geo-referenced data and is traditionally used prior to the model building phase, employing such summary measures like histograms or box plots, or three dimensional scatter diagram; (ii) *spatial statistics*, with focus on testing map patterns of phenomena (like crime of diseases), the tendency for them to cluster or disperse, and how spatially defined objects interact with one another, either statically or over time; (iii) *spatial econometrics*, which flourished after 1988 – the date of publication Anselin's influential book (under such a title) in which the well-established econometrics is related to spatial models in the form of the fundamental regression tools of the spatial econometrics, with recent extension in geographically weighted regression (GWR) to a spatial econometric system that allows regression parameters to vary over space (Fischer and Gits, *ibidem*, p. 5); (iv) *geostatistics*, defined as a way to describe and explain physical phenomena in a continuous spatial data environment (with such major topics as study of *variograms* or using *kriging*, predictive techniques of simulation applied in mostly natural resource exploration and earth science). So far, the latter are rarely used in sociological analysis, but the previous ones are being employed in analyzing spatial aspects of phenomena (ideally, using GIS data), such as inequality, power, politics, interaction and social network, community, social movements, poverty, deviance, crime, study of residential segregation, etc.

Association and causality – *counterfactual account in statistics and sociology.* Within the traditional (Hempelian) explanatory framework that has prevailed over the post-war methodology of empirical research for a while to establish causality seemed to be like finding the Holy Grail of theorization (especially within the neo-positivist perspective). However, according to Bunge (1979) quoted by Bernert - op cit.,) not all thinkers shared such a view – for instance causality has been declared a 'fetish' (Karl Pearson), or the 'relic of a bygone age‘ (Bertrand Russell), or a 'superstition' (Ludwig Wittgenstein) and a 'myth' (Stephen Toulmin), (*ibidem*, p. 231). Moreover, sociological interpretations of underlying mechanism being called into question to provide explanation in terms of causality has not generally been compatible with interpretation of association by statisticians, who typically avoid such a language. Some of them criticized causal type interpretation of earlier versions of path analysis coefficients and structural equation models often employed by sociologists (Freedman, 1987, Sobel, 1998). In brief, non-experimental, observational data which social researchers typically work with put several challenges to classic causality-based explanatory framework. On the other hand, this has inspired social science methodologists to deal with some dilemma and controversies – starting with revising some of unnecessary assumptions, such as (that) explanation is synonymous with causal explanation or that to explain also means to predict. For instance, the latter may be met by spurious causation or time series, while sociologists called for preventing against taking *Granger case* for causation (e.g., Goldthorpe, 2001) stressing that causality needs more than *robust dependency*. [So-called ‘Granger causation’ was originally discussed in the context of econometric time-series analysis (Granger, 1969) demonstrating that predictive power cannot be a criterion for causality which is meant on this ground as follows: “A variable X ‘Granger causes’ Y if, after taking into account all information apart from values of X. these values still add to one’s ability to predict future values of Y” (*ibidem*, p. 2).] And that, in addition to formal conditions – considered the “holy trinity” of causality (association, asymmetry and non-spuriousness or ‘third variable problem’, e.g. Mutz, 2011, p. 9) – should involve a type of a ‘generative process’ within a ‘substantive’ model’ that meet a ‘pragmatic utilization’ criteria while being empirically testable (*ibidem*, p. 15). Such a concern was raised in line of earlier sociologists’ caution – Bernert indicates to Znaniecki’s (1934) warning against the danger of withdrawing 'into the realm of pure mathematical concepts’ (*ibidem*, p. 233) while speaking only of association in lieu of cause (*ibidem* p. 245).

Somewhat alternative to the Goldthorp’s approach to causality though congruent with it in ignoring statistical models – including recent advances in counterfactual causal analysis (discussed below) – is a mechanism-based approach to causality being under intensive development during the past decade within *analytical sociology* (AS). It is a new approach in sociology, focused on providing explanation (in the vein of Hempel’s covering law models of explanation) “by specifying mechanisms that show how phenomena are brought about” (Hedström, 2005, p. 24). It seems worthwhile to mention about AS for its proneness to greater interdisciplinarization of social research as being open to various types of the mechanism and for its distinguishing feature which is its aspiration to provide "a syntax of explanation…(that) comprises a set of constraints on how an explanation should be constructed and empirically tested" (cf. Manzo, 2010). It is enough to confine these remarks to express the view that this approach could benefit from extending by directly embracing the counterfactual model of causality (e.g., Morgan and Winship, op cit., p. 233).

The issue of causality was more or less explicitly present in practically all the important classes of methods which were developed – some with active involvement of sociologists – during the period 1950-90, such as (e.g., Heeringa et al., 2010, op cit.,): log-linear models and related methods for contingency tables – generalized linear models, e.g. logistic regression –survival analysis model – general linear mixed models (hierarchical linear models) – structural equation models – and latent variables models. Following Lazarsfeld's 'elaboration' approach to explore relationships among the measured variables (i.e., to explain the correlation), sociologists sought to accomplish what multiple regression methods could do in experimental situations through employing path-modeling with latent variables in non-experimental analysis (i.e., using survey data in *analytical* as opposite to *descriptive* mode in Skinner's terminology – see Heeringa et al., op cit. p. 4).

During the 1970's and 1980's sociologists were refining *structural equation models* – following two paradigmatic traditions: a 'Simon-Blalock tradition' and 'Duncan-Blau' (path analysis) tradition. Further technical advancement in structural equation modeling (SEM) was due to the graphical presentation of the problem by computer scientists (Pearl, 2000). But none of them was recognized by statisticians as a satisfactory way of dealing with cause-and-effect question which motivate much of the empirical (non-experimental) work in social science. The new opportunity for quantitatively oriented social sciences emerged with moving beyond the structural equation and path-model techniques that have been prevailing during 'the age of regression' – when some of their leading proponents, notably Blalock (1964), overstressed "focus on causal laws as represented by regression equations and their coefficients' (*ibidem*, p. 177).

The move was toward experimental type of reasoning based on 'potential outcomes framework' (POM) proposed by Neyman (1923) formalized as the *counterfactual model of causal inference* (CMCI) for observational data by statisticians (Rubin, 1974 and series of works by him and his colleagues, Holland, 1986, Rosenbaum, 1989), and by econometricians (especially, works of Heckman, 1989 and Manski, 1995). The counterfactual model that brings experimental language into observational data analysis ('experimental metaphor') – in line with Stouffer's advice for sociologists to "always keep in mind the model of a controlled experiment, even if in practice we may have to deviate from an ideal model"– is now being used with increasing frequency in sociology, psychology and political science, see Morgan and Winship (op. cit., p. 4-7)". Most of the paradigmatic applications of this model which they discuss are provided by studies of cause-and-effect type of problem in the evaluation context

– for instance: (a) the causal effect of school vouchers on learning (Chubbs and Moe, 1990); (b) the causal effect of Catholic schooling on learning (Coleman and Hoffer, 1987); (c) the causal effect of manpower on earnings (LaLonda, 1995) [for instance, 'treatment' is having or not having college degree and earning (as an effect) – four theoretically possible cases/answers to 'what-if earnings' for each person, including two what-is potential outcomes which are counterfactual: adults who have completed high school, *what-if* earning under the state „having a college degree", and adults who have completed college degree, *what-if* earning under „have only a high-school degree", etc.]; (d) the causal effect of alternative voting technology on valid voting ((Wand et al., 2001).

Operationally, the essence of the counterfactual model of causality (or CMCI), a counterfactual account ("what would have been") that is built-in potential outcomes framework, refers to the relationship between potential and observed variables given the causal states, 'treatment' and 'control', respectively. Interestingly enough, the statistical models of counterfactual causation have been developed independently on counterfactual theories of causation proposed in the philosophy of science following David Lewis's (1973) conceptualization of 'world semantics' for counterfactuals ("If A had not occurred, C would not have occurred"). The central notion of this semantics is a relation of comparative similarity between possible worlds (for a pair of possible worlds, the world that resembles the actual world more than the other is said to be closer to actuality) – see Menzies (2008, p. 3). It was only after innovative re-formulation of the counterfactual causation in graph-theoretic version of the structural equation framework by Pearle (2000/2009, op cit.) when the two strands of the literature met each other (Hitchcock (2001), Woodward (2003). Some authors (e.g., Kluve, 2001) distinguish between three approaches to modeling causation for observational data: structural equation models (SEM), potential outcome model (POM), and Directed Acyclic Graphs (DAG) dominating respectively in econometrics, statistical, and formal (or computer-science) domains of applications. However, others consider these approaches as different stages in the development of structural equations for modeling causality rather than separate methodologies – see Grace et al., (2012). According to Grace et al., the first generations of SEM consists of the early works (following Wright's (1921) ideas) of path analysis that spreads to econometrics in 1940s, and to sociology (Blalock 1964), being limited however to the analysis of correlation matrices. The second generation (continuing to the present) consists of the already mentioned works of Jöreskog's (in 1970s.) synthesizing factor and path relations under the LISREL model involving both latent and observed variables. The third has just begun with contribution from computer science (Pearl 2000/2009).

The statistical counterfactual causal modeling gained recently new impetus from a graph-theoretic implementation of structural equation modeling (following ideas and methods proposed by Pearl (2000/2009)) that subsumes the historical matrix approach and is incorporated into general SEM practice (Shipley, 2009). Most characteristic of this new approach – reckoned by Grace et al., (op cit.) as third generation of structural equations, which are considered by Pearl (op. cit.) the natural language for representing and studying causal relations – is the generalization of the structural equation model as a causal graph along with extending it to the nonparametric level.

For the purpose of presenting the idea of the statistical approach to the problem – let Y^t and Y^c denote potential outcome random variables defined over all individuals in the population under study. Accordingly, Y_i^t is the potential outcome in the treatment state for individual i, and Y_i^c is the potential outcome in the control state for individual i. And the individual-level causal effect of the treatment is defined as:

$$\delta_i = Y_i^t - Y_i^c .$$

For two causal states (a binary case), a causal exposure variable D is equal 1 for members of the population who are exposed to the treatment state, and equal 0 to others (non-exposed to the treatment). Formally, the observed outcome variable Y is defined as

$$Y = Y^t \text{ if } D = 1$$

$$Y = Y^c \text{ if } D = 0$$

Or, equivalently

$$Y = DY^t + (1\text{-}D)\, Y^c$$

Consequently, in terms of the Neyman potential outcomes' conceptualization the problem –called in the literature the fundamental problem of causal inference (after Holland, op cit.) – can be presented as below:

a) Treatment Group (D = 1) & $Y = Y^t$ $\rightarrow$ Observable as Y

b) Treatment Group (D = 1) & $Y = Y^c$ $\rightarrow$ Counterfactual

c) Control Group (D = 0) & $Y = Y^t$ $\rightarrow$ Counterfactual

d) Control Group (D = 0) & $Y = Y^c$ $\rightarrow$ Observable as Y

Since, in reality, the potential outcome under the treatment state can never be observed for those observed in the control state, and *vice-versa* (rows b and c), the outcome variables (such as labor market earning, or test score, or voting results in the above cited examples) contain only a portion of the information needed to calculate individual-level causal effects. [This is why, following Rubin (1978), counterfactual causal analysis at its core is considered a missing data problem (Winship and Sobel, 2001, p. 14).] As a consequence of impossibility to calculate individual-level causal effects, the average treatment effect (ATE) in the population is estimated as the expectation, E(.) of a difference between potential outcomes:

$$\begin{aligned} E[\delta] &= E[Y^t - Y^c] \\ &= E[Y^t] - E[Y^c] \end{aligned}$$

For instance, the average causal effect in the case of study by Coleman et al. (op cit.) is then estimated as the mean value among all students in the population of these what-if differences in test score (for a randomly selected students from the population). Often, the average treatment effect is defined separately – as conditional average treatment effect – for those who typically take the treatment, ATT (assuming the linearity of the expectation operator)

$$\begin{aligned} E[\delta \mid D=1] &= E\,[Y^t - Y^c \mid D=1] \\ &= E[\quad Y^t \mid D=1] - E[Y^c \mid D=1] \end{aligned}$$

and for those who typically do not take the treatment, ATC

$$\begin{aligned} E[\delta \mid D=0] &= E\,[Y^t - Y^c \mid D=0] \\ &= E[\quad Y^t \mid D=0] - E[Y^c \mid D=0] \end{aligned}$$

Framing the experimental type of reasoning (causal inference) for observational data through employing counterfactual models in social science has resulted in elaborating analytical instruments capable to compensate for inherently deficient data (POM as ‘missing data problem’). The matching estimators that are the classic technique for estimating causal effects are typically interpreted either (a) as a method to form quasi-experimental contrast by sampling comparable units from a population under study, or (b) as a nonparametric method of adjustment for treatment in the situation when parametric regression is not persuaded. The significant use of this technique in sociology was as early as mid-1980s. (e.g. by Berk and Newton, 1985). Different procedures of matching estimators are under continuing refinement, such as: matching as conditioning via stratification – matching as weighting (using propensity score as the estimated probability of taking the treatment as a function of variables that predict treatment assignment) – matching as a data analysis algorithm – matching when treatment assignment is non-ignorable (Morgan and Harding, 2006). However, none of these methods is considered perfect, being either limited in practical applications (e.g., only recently included are multivalued treatments to procedures being essentially appropriate for binary treatments or exposures). Also, predicting treatment status from the observed variables with a logit model may not solve the causal inference problem (e.g., if no variable yielding a ‘perfect stratification’ is known).

In their excellent exposition of methodological issues involved in counterfactual causal modeling, Morgan and Winship (2007/2009) discuss the prospects and limitations of this approach in empirical research in social sciences. At first, they stress its utility *vis-à-vis* traditional structural equation modeling pointing first to data related advantages (such as sufficiency of the requirement that data be balanced with respect to the determinants of treatment assignment, while easing the problem of omitted variables). On the other hand, they specify some objections to employing counterfactual modeling, such as (1) incapability to non-manipulable causes, (2) inappropriate for discovering the causes of effects, and (3) causal inference should not depend on ‘metaphysical quantities’ (*ibidem*, p.278-282).

They offer four modes of causal inquiry in observational social science which constitute a sequence of cumulative proprieties: (a) *associational analysis* – association as a precondition of causation; (b) *conditional associational analysis* – to eliminate sources of spuriousness; (c) *mechanism-based analysis* – efforts to provide mechanistic explanation of the process that generates the causal effect; and (d) *all-cause structural analysis* – complete set of variables between the putative causal variable and the outcome variable (the structural equation models in economics provide an exemplary successful usage of such a mode of causal analysis). Of course, the last approach that could be seen as an ultimate goal of causal inquiry – attempting to provide answer to all of the “who, when, where, and how” questions – calls for a theory and data, as any type of causal approaches does. In this point, however, counterfactual models of causation are useful for identifying in a more accurate way which stage (or mode) of causal analysis is actually feasible, given a theory and data availability.

As regards the problem of data that strongly interferes with making a choice among feasible strategies of causal analysis, a final observation relates to an alternative approach to causality – which could be considered as data-based pragmatic approach due to actually abandoning the problem of causality and interpretation of the nature of cause-and-effect relationship. Namely, it is a newly proposed data collection technique which under the name *population-based survey experiments* uses “survey sampling methods to produce a collection of experimental subjects that is representative of the target population of interest for a particular theory” and is easily implemented through utilizing computer-assisted telephone interviewing (CATI) and internet-based interview (Mutz, 2011, p. 2-10). The underlying idea is to enable the researchers to control the random assignment of participants to variations of the independent variable in order to observe their effect on a dependent variable. Their proponents, who tested efficiency of such quasi-experimental reasoning based on sample survey in Time-sharing Experiments for the Social Science (TESS) while stressing internal validity of the approach as its strength are, however, aware of the problem of external validity and generalizability as its fragile part, that needs further elaboration. (see Mutz op cit., p. 22). Anyway, it illustrates another attempt of sociologists and other social scientists to contribute to the statistical methods in the difficult context of establishing empirically unbiased causal inferences. [It seems promising in the theory-based evaluation research context, albeit not discussed in the literature yet.]

III. Conclusions

The mutually supportive interplay between statistics and sociology over the post-war period that contributed to the developments in each of these disciplines substantially has become an object of systematic reflection (starting as early as with Lazarsfeld's historical notes on quantitative sociology in late 1950s and with Bernert's paper at early 1980s of the 20th century – not to mention several important papers written by statisticians, such as by Kish on sampling or Stigler and Fienberg and others on different issues). The confluence could be observed in many areas, especially of those being motivated by going beyond descriptive use of statistics and searching for an explanation – from Lazarsfeld's 'elaboration' program and Goodman's log-linear models through logistic regression to path analysis and structural equation models of causation (Blalock, Duncan, others) to recent counterfactual causal modeling. A lot of new hopes have been created with introduction of the counterfactual model of causal inference for observational data due to, in a summary: (i) the possibility that the size of a treatment effect may vary across individuals (*effect heterogeneity*); (ii) the researcher need to fully specify the *implicit manipulation* or "experiment" associated with the estimation of a causal effect ('no causation without manipulation' as emphasized by Holland and Rubin); (iii) selection of variables as controls given that they substantially change the estimate of the treatment effect; (iv) using *matching* method due to its ability to provide a powerful nonparametric alternative to regression (for the estimation of a causal effect); (v) limiting *instrumental variable* estimators only to estimating the effect of the treatment for those individuals whose treatment status is changed; and (vi) longitudinal data do not suffice for causal inference without prior assumptions about the values of the outcome (under the counterfactual condition).

One of the important lesson learnt from such studies is the need for a more interdisciplinary orientation than it was during 20th century in quantitative methods, marked by isolation from each other (and often from statistics as a whole, as pointed by the critics of structural equation models with latent variables in 1960-70s). Interdisciplinarization is predicted (and advocated) by leading social methodologists and is stressed as chief direction of developments in quantitative sociology. Importance of such an approach has been recognized in contemporary institutional efforts towards *interdisciplinarization* of social science research through establishing several joint programs of research and education, both by departments of statistics and by several social science departments in Europe and the USA. In the context of the Congress of the Polish Statistical Association, it seems worthwhile mentioning that such a kind of institutional interdisciplinary cooperation – sociology and statistics – was created in the late 1960s., along with establishing of a Statistical-Sociological Research Unit at the Central Statistical Office of Poland[1].

[1] The Unit was created under leadership of Aleksander Wallis and resumed (after his depart) by Krzysztof Zagórski. (The author was a part of the 'first generation' Unit's members).

REFERENCES

AGRESTI A., 1990. Categorical Data Analysis. Wiley. N. Y.

ALLISON P., 1982. Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology,* 13, pp. 61-98.

ALONSO, W., STARR, P., 1987. The Politics of Numbers, Russell Sage Foundation, NY.

BARTHOLOMEW, D. J., STEELE F., MOUSTAKI I., and GALBRAITH J. I., 2002. The Analysis and Interpretation of Multivariate Data for Social Scientists, Chapman and Hall/CRC, Boca Raton, Fl.

BARTHOLOMEW, D. J., KNOTT M., MOUSTAKI I., 2011 Latent Variable Models and Factor Analysis: A Unified Approach. 3rd edition; *Wiley Series in Probability and Statistics*. John Wiley and Sons, Ltd.

BERNERT, CH., 1983. The Career of Causal Analysis in American Sociology, *The British Journal of Sociology,* Volume 34 Number 2 (June).

BLALOCK, H. M., 1961. Causal Inferences in Nonexperimental Research. New York: W.W. Norton.

BLAU, P. M., DUNCAN O. D., 1967. American Occupational Structure. Free Press, N.Y.

BUNGE, M. A., 1979. Causality and Modern Science. (3rd ed.), New York: Dover.

Camargo Alexandre de Paiva Rio, Sociology of Statistics: possibilities of a new field of investigation, Hist. cienc. saude-Manguinhos, vol.16 no.4 Rio de Janeiro Oct./Dec. 2009. (http://dx.doi.org/10.1590/S0104-59702009000400004)

CAMIC, CH., XIE, Y., 1994. The Statistical Turn in American Social Science: Columbia University, 1890 to 1915. *American Sociological Review,* Volume: 59, Issue: 5.

COX, D. R., 1971. Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 34, pp. 187-220.

DESROSIERÈS, A., 2011. Words and Numbers. For a Sociology of the Statistical Argument, chapt. 2 [in] Saetnan A. R., Lomell H. M., Hammer S. (ed.): The Mutual Construction of Statistics and Society, Routlege, New-York, pp. 41-63.

DUNCAN, O. D., 1979. How Destination Depends on Origin in the Occupational Mobility Table. *American Journal of Sociology* 84:793-803**.**

FIENBERG, S. E., A Brief History of Statistics in Three and One-Half Chapters: A Review Essay, *Statistical Science*, Vol. 7, No. 2 (1992), pp. 208-225.

FISCHER, M. M., GETIS, A., 2010, Handbook of Applied Spatial Analysis. Software Tools, Methods and Applications, Springer, Berlin Heidelberg.

FISHER, R. A., 1935, The Design of Experiments. Oliver and Boyd, Edinburgh.

FREEDMAN, D. A., 1987. As Others See Us: A Case Study in Path Analysis. *Journal of Educational Statistics,* 12: 101-223.

GOLDTHORPE, J., 2001. Causation, Statistics, and Sociology, *European Sociological Review*, Vol. 17 No. 1, 1-20.

GOODCHILD, M. F., ANSELIN, L., APPELBAUM, R. P., HARTHORN, B. H., 2000. Toward Spatially Integrated Social Science, *International Regional Science Review* 23, pp. 139-159, (April).

GOODMAN, L. A., 1979. Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories, *Journal of the American Statistical Association* 74:537-52.

GOODMAN, L. A., 1985. The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories**.** *Annals of Statistics* 13:10-69.

GRACE, J. B., SCHOOLMASTER jr., D. R., GUNTENSPERGEN, G. R., LITTLE A. M., Mitchell B. R., Miller K. M., Schweiger E. W, 2012. Guidelines for a graph-theoretic implementation of structural equation modeling, *Ecosphere* 3(8):73.

HANSEN, M. H., MADOW, W. G., TEPPING, B. J.(ed.), 1983. An evaluation of model-dependent and probability-sampling inference in sample surveys, *Journal of the American Statistical Association*, 78: 776-793.

HECKMAN, J. J., 1979. Sample Selection Bias as a Specification Error, *Econometrica*, 47:153-61.

HEERINGA, S. G., WEST B. T., BERGLUND P. A., 2010. Applied Survey Data Analysis, Chapman and Hall/CRC, Boca Raton, FL.

HITCHCOCK, C., 2001. The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*, 98: 237-299.

HOUSE, J. S., JUSTER, F. T., KAHN, R. L., SCHUMAN, H., and SINGER, E. (ed.), 2004. A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond, University of Michigan Press.

JORESKÖG, K. G., 1973. A General Method for Estimating a Linear Structural Equa-tion System. Pp. 85-112 [in] A. S. Goldberger and O. D. Duncan (ed.) *Structural Equation Models in the Social Sciences*, Seminar. N. Y.

KISH, L., 1995. The Hundred Years' Wars of Survey Sampling, chap. 3 [in] Graham Kalton and Steven Heeringa (ed.), *Leslie Kish: Selected Papers*, 2003. John Wiley & Sons, Hoboken, NJ.

KLUVE, J., 2001. On the Role of Counterfactuals in Inferring Causal Effects of Treatments. Alfred Weber Institute, University of Heidelberg and IZA Bonn Discussion Paper No. 354.

LAZARSFELD, P. F., 1961. Notes on the History of Quantification in Sociology--Trends, Sources and Problems. *Isis*, Vol. 52, No. 2. (Jun., 1961), pp. 277-333.

LAZARSFELD, P. F., 1950."The Logical and Mathematical Foundation of Latent Structure Analysis. p. 362-412 [in] *Studies in Social Psychology in World War II. Vol. 4, Measurement and Prediction*, edited by E. A. Schulman, P. F. Lazarsfeld, S. A. Starr, and J. A. Clausen. Princeton, NJ: Princeton University Press.

MANZO, G., 2010. Analytical Sociology and Its Critics, *Archives of European Sociology*. LI, 1 pp. 129–170.

MENZIES, P., 2008. Counterfactual Theories of Causation. *The Stanford Encyclopedia of Philosophy*, Vol: 2011, Issue: Dec. 16, Stanford University.

MORGAN, S. L., HARDING, D. J., 2006. Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. *Sociological Methods and Research* 35: 3-60.

MORGAN, S. L., WINSHIP, C., 2007. Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press, N.Y.

MUTZ, D. C., 2011. Population-Based Survey Experiments, Princeton University Press, Princeton, NJ.

NEYMAN, J., 1933. An outline of the theory and practice of the representative methods: The method of stratified sampling and the method of purposive selection (in Polish – English version published in 1934. On the two different aspects of the representative method. *Journal of the Royal Statistical Society*, 97: 558-625).

OKRASA, W., 1999. The Dynamics of Poverty and the Effectiveness of Poland's Safety Net. *Policy Research Working* Paper No 2221. The World Bank, Washington D. C.

OKRASA, W., 1999. Who Avoids and Who Escapes from Poverty during the Transition. Evidence from Polish Panel 1993-96, *Policy Research Working Paper* No 2218. The World Bank, Washington D. C.

OKRASA, W., 2012. Spatially Integrated Social Research and Official Statistics: Methodological remarks and empirical results on local development, Paper presented at the *International Conference of Spatial Econometrics and Regional Economic Analysis*, Łódź, 4-5 June (2012).

PEARL, J., 2009. Causality: Models, Reasoning and Inference. (2nd edition), Cambridge University Press, N. Y.

PEARL J., 2012. The causal foundations of structural equation modeling. Pages 68–91 in R. H. Hoyle (ed.). *Handbook of structural equation modeling*. Guilford Press, New York, New York, USA.

RAFTERY, A. E., 2001. Statistics in Sociology, 1950-2000: A Selective Review. *Sociological Methodology*, Vol. 31, pp. 1-45 American Sociological Association.

SHIPLEY, B., 2009. Confirmatory path analysis in a generalized multilevel context, *Ecology* 90:363–368.

STIGLER, S. M. (1986), The History of Statistics: The Measurement of Uncertainty before 1900, Harvard Univ. Press, p. 410 [Reissued in paperback edition (1990)].

SOBEL, M. E., 1995. Causal inference in the social and behavioral sciences, [in] G. Arminger, C. C. Clogg, and M. E. Sobel ed. *Handbook of Statistical Modeling for the Social and Behavioral Sciences,*Plenum Press, N. Y.

SOBEL, M. E., 1998. Causal Inference in Statistical Models of the Process of Socio-economic Achievement: A Case Study. *Sociological Methods and Research* 27: 318-48.

STARR, P., 1987. The Sociology of Official Statistics [in] Alonso, W., Starr, P., (ed.). 1987. *The Politics of Numbers, Russell Sage Foundation*, N. Y., pp.7–58.

Statisticians in History. http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=11.

URRY, J., 2004, The Sociology of Space and Place, [in] Judith R. Blau (ed.) *The Blackwell Companion to Sociology*. Blackwell Publishing, Malden, MA.

WILCOX, W. F., 1930. Census, *Encyclopedia of the Social Sciences*, New York: Macmillan.

WOODWARD, J., 2003. Making Things Happen. A Theory of Causal Explanation. Oxford University Press, Oxford.

ZNANIECKI, F., The Method of Sociology, Farrar & Rinehart, New York 1934.