UNIVERSITÀ DI PARMA

UNIVERSITA' DEGLI STUDI DI PARMA

DOTTORATO DI RICERCA IN
"INGEGNERIA INDUSTRIALE"

CICLO XXIX

An interactive MATLAB device for chi-squared tests about Markov Chains and estimating nine transition-matrix-based mobility indices: evidences from AIDA microdata of some Italian industries.

Coordinatore:
Chiar.mo Prof. Gianni Royer Carfagni

Tutore:
Chiar.mo Prof. Piero Ganugi

Dottorando: Dott. Danilo Aringhieri

Anni 2014/2019

To my parents and to my brother

Index

Appendix A3. Estimated values of the nine transition-matrix-based mobility indices in MarkovInfer_MobInd.m, computed for the 1st-order stationary Markov Chain and for the not-stationary one, both in the date-by-date version as and in the aggregated-in-time variant, for every one of the 28 combinations [industry; economic variable].

Abstract

The present doctoral dissertation plays on discrete Markov Chains and hypothesis statistical assessments about them; on inferential estimation from microdata, mobility indices and Matlab programming. The chains are discrete because the structure of the underpinning state-space is discrete (i.e. it is parcelled out into a finite number of limited classes, except the upset one), as well as the temporal parameter *t* is not-continuous and discretely-marked, to be coherent with the cadence of registered observations in databases for Economics and Social Sciences. Hypothesis tests avail of the chi-squared distribution and derive from the Maximum Likelihood Principle; any other mathematical principle, potentially helpful for developing inferential techniques, like the Minimal Entropy Principle (where the entropy is the Shannon informational one) is not employed. We shall not use neither census data nor time series; whatever the historical lapse fixed by the data and whatever the data sample, we wanted to target the largest core of economic agents persisting for all observations spread out on that epoch. Thus, the concept of mobility inside a population *will have always to be intended here in its intragenerational sense* and it has been watched through the lens of the transition-matrix-based approach. Shades of the mobility have been connotated via an ensemble of indices defined in force of the transition matrix itself; its transition probabilities are estimated for the data in the case of a first-order Markov chain, thence those indices are discrete respect to the time parameter, too. Everything was transposed into the programming language for the Matlab environment.

In most areas of the United States since the years of World War II, rapid changes in the number and size distribution of dairy, crop and livestock farms seemed to occur (for example, in the twenty-year periods between 1958 and 1977, as stated in Stavins and Stanton (1980), and from 1969 to 1982, as stated in Disney et Alii (1988), among others), while population was growing. Fearing in the 1950s the U.S. government that either underproduction could afflict the population already stressed by the War, or overproduction could cause the plunge of prices (thence, the failure of many firms), in order to steer

regulatory policies, across the last decades applied statisticians have intensively utilized manageable Markov models to examine changes in businesses size distributions and to project such changes toward future ages. Meanwhile, politicians and qualified observers began to suspect some serious phenomena of concentration were occurring in more than one american industry. Moreover, social scientists had discovered that transition matrices, estimated for a first-order Markov chain, could have been an instrument to capture people's movements between income strata and offspring's changes of social location compared to their fathers' classes (Prais 1955). From then on, those topics have not yet lost their allure in Economics above all, even if not exclusively, in order to yield future projections advertising policy making involving enterprises and consumers. The modern versatility of manifold Markov models is well illustrated in the monograph Ching, Wang, Ng, Siu (2013).

The dissertation's first part describes a Matlab executable script (with extension *.m), called "*MarkovInfer_MobInd.m*", which implements, after automatically importing and 'softly cleaning' any xls or xlsx-type spreadsheet where properly formatted data are registered, some features of the statistical inference about Markov Chains, and computes nine mobility indices defined as functionals of the transition matrix discussed in the scientific literature. Moreover, this tool automatically brings about two different χ^2-type tests for the same chains on the referring scientific database, AIDA, commissioned by the Union of the Italian Chambers of Commerce to the company 'Bureau Van Dijk' about all balance-sheet variables of the Italian firms. Although its time series are very short (only ten years), the AIDA archive has the merit of containing almost-complete firms' samples respect to a predetermined set of quantitative or qualitative items, which are highly representative of the Italian businesses' true sub-population under those items, enabling reliable statistical examinations.

The mobility measures (in a purely intragenerational acceptation), based on the transition matrix, are the Normalized Directional Index by Ferretti and Ganugi, the Alcalde-Unzu Normalized family of Relative Indices of mobility as a movement; the Trace Index and the Exponentiated (or Generalized) Determinant one; the Second-Eigenvalue Mobility as well as the All-Eigenvalues one; the Index of Predictability, the 2nd Bartholomew Index and,

finally, the period-invariant asymptotic half-life Index. At first, our code extracts from the imported data the fixed subsample of all agents exhibiting a value for the studied variable, for each of the dates registered in the spreadsheet. Then, all the mobility measures, both in the one-period (or step-by-step, or date-by-date) version and in the compound form pertaining to aggregated observation periods, are calculated in conformity with the two simplest Markov models: the 1st-order time-homogeneous chain and the 1st-order stationary one. Such two models are also used by *MarkovInfer_MobInd.m* to estimate the mean permanence times inside every state of the chains. Mean survival durations within every class are then counted through a third, purely mechanistic methodology, without relating to any theoretical background.

Two diverse couples of statements, in the form [null hypothesis versus alternative one], indicated as '*antinomies*' too, are verified both by the chi-squared test, derived from the homogeneity test for contingency tables, and by the maximum likelihood ratio criterion (MLRC), which are suggested in the 1957 paper by Anderson and Goodman. The null-alternative statements are: the first-order and time-homogeneous Markov chain against the first-order and not-stationary chain (*Query a*), along with the first-order stationary Markov chain against the second-order stationary one (*Query b*). The number of degrees of freedom is not the maximal nominal value (defined as a product of the basic parameters of the Markov ensemble of states) but it is sensibly reduced taking into consideration some problems (like divergence phenomena of the kind $n/0$, as well as 0/0 indeterminate forms of the empirical statistics) that may materialise when an hypothesis's trustworthiness is explored onto actual data, after having been conceptually defined. It is a lesson outlined in Bickenbach and Bode (2003).

Our device repeatedly interacts with the user via a command line interface (CLI). It displays some helpful captions and some fundamental sample estimators, e.g. the array of selected quantiles, to guide and advise him; then, it demands to enter as inputs the preferred values and shapes for all the indispensable 'free parameters' specifying the Markov models and the mobility measures. For instance, during every run of the program the user must point out the set of classes underlying the chains, inserting their number together

with the inner extremes of the partition of the variable's domain which delimit the pairwise-disjoint bands.

At the end of the run, all the quantitative outputs (including the results of chi-squared and MLR hypotheses tests) are saved into the Matlab workspace as bidimensional or multidimensional arrays, rather than as vectors or scalars. Some graphical outputs are also produced concerning the mobility indices on the two first-order Markov models, to make a comparative analysis feasible among these indices and to make the study of their dependence on time possible. If one economic variable is chosen and its data are processed for disparate industries, the researcher could think over how either hypothesis-tests' outcomes or mobilities' patterns change when the industrial sector has been changed.

The second part of the dissertation is dedicated to the presentation and interpretation of some results from the data elaboration via *MarkovInfer_MobInd.m* and focuses on five Italian manufactures, classified into the ATECO 2007 taxonomy, year by year from 2006 until 2015, at the national scale (corresponding to the first two digits, after the letter 'C' for any manufactural sector, inside the complete script), consisting of thousands and thousands of agents: Food and Beverage Industry (C10), Rubber and Plastic Products (C22), Textiles (C13), Machinery and Equipment Industry (C28), Chemicals (C20). In addition, data pertaining to two other firm groups, not expressly included in the ATECO classification but separately published in the AIDA archive anyway, have been processed: Italian Small and Medium Innovative Enterprises (or PMI Innovative, also indicated as Inno-SMEs in the remainder of the thesis), composed of just over 300 businesses, and Italian Innovative Start-Ups. For every industry or group four balance-sheet variable has been evaluated: Revenues from Sales and Services; Total Assets; Total Payables; Total Production Monetary Value. So, official data were downloaded for 28 combinations of the kind [Manufacture, Economic Variable], obtaining 28 corresponding Excel single spreadsheets, formed by all companies lasting within the sheet for the entire decade 2006-2015, which were worked one by one through our Matlab tool to produce 28 Matlab workspaces. Each workspace is full of all outputs, both the conclusive outputs and the intermediate ancillary ones, computed by the code, identified by their own names, and stored as objects of many varied dimensions: some of them are

scalars or vectors, some are matrices (that is bidimensional arrays), some others are multidimensional arrays. Owing to the exiguous number of units (12-14) enduring in the archive during the decade, elaborated outcomes for the group of Innovative Start-ups have not any scientific reliability and such group is in practice only an example of misleading inferential estimations. For this reason, Innovative Start-ups will not be exhaustively commented.
About tests, bolstered by the asymptotic chi-squared distribution, on the two previous antinomies, in *Query b* the stationary first-order is defeated by the stationary second-order, for all the four variables of the five ATECO industries and the Inno-SMEs group, and occurrence probabilities are infinitesimal. Instead, in *Query a*, the stationary chain is generally rejected versus the not-stationary one for the five ATECO divisions, by infinitesimal *p*-values, but the stationary first-order appears to become more credible in the case of Innovative SMEs.
Various regularities, and some unexpected facts, clearly have emerged, from the analysis of the great amount of elaborations spread over the 28 workspaces, for the nine mobility indices, calculated via our program.

The thesis is formatted as follows. After the Abstract, Chapter 1 and Chapter 2 attempt to recapitulate some landmarks in the history of Markov Chains expended in Socio-Economic Sciences, with an emphasis on researches involving microdata and time series over which χ^2-type tests are exploited (only few capital works established on macro-data are overviewed). Chapter 1 tells about their ascent and successes as the most appreciated stochastic theory for Economics, Chapter 2 accounts for the efforts to enhance forecasting ability of the Markov paradigm when facing more and more challenging datasets of increasing size. Chapter 3 introduces the organisation, traits and aims of the Matlab executable file *MarkovInfer_MobInd.m*, elucidating in its last paragraph some possible investigations to be done using the outputs of the device. There would have been no time enough to comment in detail all the device's code, therefore in Chapter 4 and Chapter 5 we have paid attention only to some aspects, which are not obvious translations in Matlab of specific mathematical equations, of the first half of *MarkovInfer_MobInd.m*. Chapter 4 treats the automatic import of data, the simultaneous 'soft cleaning' and some of the associated problems we faced, then it instances how some interactions

with the user have been designed. In Chapter 5 the construction of the Markov state space, starting from micro-data, is explained and the control on the mathematical correctness of bounds, inserted by the user to split the space into classes, is illustrated. In its first part Chapter 6 introduces and comments the sets of individual data from AIDA (the microdata, or 'raw data', or granular data), to be processed by the tool, and the configuration of the indispensable 'free parameters'; in the second part a list of the principal outputs, saved into the Matlab workspaces, is reported indicating their name and revealing sense and dimensions of each of them. Chapter 7 and Chapter 8 tell about some data-analysis questions which can be tackled via our software. In Chapter 7, the results of χ^2-type tests about Markov Chains in Economics from our bibliographical references are recalled; thereafter, general outcomes of both the Pearson's test and the maximum likelihood ratio criterium (MLRC), from our tool on AIDA raw data, are tabulated and discussed regarding the two antinomies cited above. Chapter 8 is devoted to the analysis of the great amount of computations about the nine transition-matrix-based mobility indices, bringing to light evidences and regularities dealing with those measures in the selected Italian ATECO sectors. The recapitulation of the principal discussions and results, proposed throughout the thesis, are provided in the Conclusion. After the Bibliography and Webography, the text of the code *MarkovInfer_MobInd.m* is copied inside the first Appendix A1. The following Appendix A2 contains tabulations filled with numerical details of the Pearson's test and MLRC limited to the Foods and Beverages national ATECO division and to the group of Innovative Small and Medium Enterprises, as an example of what can be learnt about Markov Chains from any one of the 28 Matlab workspaces we have produced. The ending Appendix A3 displays in tables all the computed values of the nine mobility proxies, in the mono-period form (section A3.1) as well as in the time-aggregated one (section A3.2), estimated through both the 1st-order chains, for the 28 combinations [balance-sheet item; industry] we looked at.

***The doctoral candidate Danilo Aringhieri is the sole responsible maker and author of the whole work, summarised in this abstract*.**

Chapter 1
Historical synopsis of Markov Chains in Economics: the ascent

Since when Vilfredo Pareto demonstrated, in his 1897 "Cours d'économie politique", that the allocation of wealth and the distribution of the values of the income variable in a population are well reproduced by the statistical law bringing his name, it was evident that the nature of economic variables must be random rather than deterministic and so their dynamics over time must be necessarily explained by stochastic processes.

Obviously, we do not pretend in this single chapter to account accurately for all progressions of stochastic processes' applications to economics, in general, and their successfulness from an historical point of view.

We would rather outline a telling historical summary, restricted to our bibliography, about the confidence initially enjoyed by simple stationary Markov chains to render the dynamics of the most important socioeconomic quantities and, then, about the subsequent criticisms on their dependability. However, the power of the analysis based on transition matrices is recognized still today in statistical economics, where the attention to conditional transition probabilities fostered a further approach to the complex topic of mobility measures (to have an idea, see the fifth paragraph in the second section and the fifth section in Fields and Ok (1999)).

1.1 Some keynotes about data structures

Let us remind any researcher in Economics and Social Sciences may meet four fundamental typologies of statistical data: *census data*, *cross-sectional dataset*, *time series*, *longitudinal panel data*. The clearest forms of these structure account for one fixed variable, either qualitative or a quantitative one. In a *census dataset* (or *macro-data set* or a set of *aggregated data*) it is reported at a certain date the number of agents staying within every one of the intervals (or categories) dividing the domain of the variable, without making identities of

single agents explicit. Any data array, where all single operative units are specified one by one, is indicated as a *microdata set* (or *set of granular/raw data*). A *simple* (*or strictly*) *cross-sectional dataset* consists of a detailed list of individuals for every one of which is registered the value of the variable in a precise point of the time. It is possible to conceive the cross-sectional format in a broader sense (or better, on a broader size) in the case all statistical units are scattered on a more, or less, long-lasting period which is the primary measure of time, even if it may be articulated, in its turn, in two or more dates, considered not important now from a statistical point of view. Such a configuration is a *time-extended* (or *time-expanded*) *cross-sectional panel*; although some units may recur inside it at different dates, it should be better these recurring individuals are as less as possible (negligible, if possible), with the aim of obtaining credible results from any mathematical analysis. We are in front of a *simple time series*, instead, if the values of the variable are known for one, and only one, individual (the total output from a farm, the price of a stock) on a string of observation dates (one year after year during many decades, one day after day during many years, one millisecond after another one on a period of many months) marking the passage of time (the longer the string is, the better it is in an inferential perspective). When the variable is repeatedly assessed in a list of individuals for a sequence of observation instants, such microdata can be interpreted either as a series of many simple cross-sectional datasets or, equivalently, as an overlapping of many distinct time series over each other. It is a *longitudinal panel*, which can also be intended as a *time-extended cross-sectional panel*, where many statistical units are recurring frequently, ergo affected by 'some degrees of counterfeit'. A longitudinal panel is *balanced* if the specific response respect to the variable is available at each date for every agent comprised within microdata. Otherwise, the panel is *unbalanced*, where some units lack of the datum at some dates. The two alternative ways of defining a longitudinal panel make us understand we can generally experience interweaving of different data structures in statistical economics: e.g., timeseries might be registered for an economic proxy not for a single individual but for the aggregated value (or the mean value) of the proxy over an extended sample; very often census distributions of a country are provided at different years, owing to the periodic census survey of its nation, drafting a short time series of distributional classes.

1.2 Début of Markov Chains in Economics

Since the 1930s, Champernowne was one of the first scholars who organically investigated not only static measures of inequality implicit inside incomes frequency distributions, but also modifications in the positions of individuals among distributional strata in a certain historical period by some supporting aleatory models, based on an 'enumerable-infinite movement matrix' (which often remains constant through time). In addition, he gave for granted some variants of the law of proportionate effects on earnings. These models, where, to be honest, the assumption of enumerable infinity for the set of states is not very realistic, are exposed in his fellowship dissertation "The Distribution of Income between Persons" (deposited at first in the Library of King's College at Cambridge in 1937 and ultimately published in 1973, after some stages of revision) whose ideas would have been put into practice, for instance, in his 1953 paper "A Model of Income Distribution". In this way, a very extensive workstream was opened in statistical economics.

On such a wake, few years later, Prais tried to evaluate intergenerational social mobility in Britain and obtain an equilibrium distribution by the theory of the *Markov chains* (Prais (1955)), which at that time had already been spent in a number of fields of natural sciences and had been comprehensively expounded in Feller (1950) (a meaningful reference up to date for the stochastic calculus). He derived the distributions and the *transition matrix*, belonging to a *discrete-time, first-order, stationary Markov chain* based on seven states, from the statistical interchange tables edited by D.V. Glass in 1954. These tables illustrated the relation between the social positions of fathers and the positions of their adult sons as derived from interviews by the Social Survey in 1949 involving a random panel of 3500 males, aged 18 years and over, resident in England and Wales. The elements of such a finite-dimensional matrix were regarded as the empirical estimations for the constant-in-time conditional probabilities of families' movements from a social status to another one during a generation, whatever its position in time (stationarity). The second assumption of the Prais analysis is that the influence of any ancestors in determining sons' status is entirely transmitted only through fathers (first order chain). It was a pioneering dissertation on intergenerational social mobility. According to first-order, stationary Markov chains, the average time spent in

each class is computed. Lastly, in this work there is an early attempt to study quantitatively the variations of social mobility with time by two concepts. The first is the definition of the *perfect mobility's condition* as independence, from the placing of the progenitor, of entrance probabilities of the offspring in any social category. The second is an immobility ratio defined as the ratio between the coefficients inside the principal diagonal of the transition matrix and the components of incomes limiting distribution. We point out in advance that, ***throughout the present work, we shall overlook evolution among classes in an intergenerational sense, to focus on intragenerational migrations, from a state to another one, involving not persons but firms***.

In the 1950s, besides the study of individuals movements inside earnings distributions and intergenerational social mobility, the Markov approach got the attention of researchers involved in the topics of industrial organization. In Hart and Prais (1956) business concentration was inspected by analysing matrices composed by *conditional transition probabilities*, based on size intervals for firms in the British industry, without any derivation of an equilibrium manufacturing configuration, owing to some problems, affecting the original data, in realistically capturing entry-and-exit phenomena of firms respect to the industry itself.

In Adelman (1958) a reasonable (for the Author) long-run size distribution, based on the idea of dynamic equilibrium, was proposed for the variable of annual total assets for steel industry's companies in the U.S.A. from 1929 to 1939 and, after the pause for the World War II, from 1946 until 1956. It was assumed, in the double seven-state space (one space for each decade) holding up the first-order stationary Markov model, the existence of a *zero*th size class (or input-output reservoir) containing both the potential entrants and all the spilled-out agents. It was also proposed an index of corporate mobility at an arbitrary time, hinged upon the notion of mean lifetime inside a certain state, in case of a perfect-mobile industry defined via the independence of finals distributions from the starting state, similarly to the condition of social intergenerational perfection *à la Prais*. Data were extracted from the Moody's Manual of Investment, Industrial Securities (New York, 1930-1957). In the stationary Markov layout, integrated by the input-output reservoir, to account for entry into as well as departures from the industry, Adelman stated to have found out "a tendency towards de-concentration, as well as a growth in the size

of the median firm, [that are] trends [already] forecast on other grounds, [so] it would appear that […] the technique presented does not lead to absurd results and, therefore, that it may prove quite useful in the study of industrial structure" (within any pair of square brackets, integrations by the doctoral candidate are reported to improve the citation's comprehension). Apropos of the adjustment of the input-output depot, which would have been often employed in many researches of the succeeding decades, even if it makes the Markov model as "dynamic" in an elementary way, it has also some hidden flaws revealed in Stanton and Kettunen (1967) (see the Paragraph 3.3 for more details).

Growing popularity of Markov processes both in the natural sciences and in economics encouraged great statisticians, during the same 1950s, to develop rigorous inferential procedures about them, like the studies in Anderson and Goodman (1957) from the classic (namely, not-Bayesian) point of view. They considered statistical inference methods for Markov chains of different orders when there are many observations in every initial state, even if the panel size remains fixed as time goes by. They displayed maximum likelihood estimates for transition probabilities, with their asymptotic distribution, when the number of observations increases and the order of the chain is arbitrary. They took also advantage of likelihood ratio tests and chi-square tests, in the version already known for contingency tables, to verify several hypotheses including whether first-order chain's transition probabilities are constant or whether the Markov process is an arbitrary uth-order one against the alternative it is rth-order but not uth. Curiously, until the beginning of the 21st century these tests were poorly used in mathematical researches for economic sciences. Some exceptions (see Chapter 2, too) before Bickenbach and Bode (2003) are, to our knowledge, Judge and Swanson (1962), Hallberg (1969), Duncan-Lin (1972); indirectly, also Shorrocks (1976), Schluter (1997) and, to unveil asymmetries in the one-step incremental process extracted from time series, Neftci (1984) together with Nelson, Braden, Roh (1989). The same tests are the subject of the first part of our data analysis (Chapter 7).

1.3 Triumph of the Markov first-order stationarity in Economics

The lessons in Prais (1955), Hart and Prais (1956), Adelman (1958) inspired a great confidence (not always statistically assayed relative to any sample of economic agents) in spending Markov theory to gauge future distributions, business concentration and turnover: their strategies and techniques were reproduced, for instance, in two 1961's articles by Collins and Preston, and in the 1962's work by Judge and Swanson.

In Collins and Preston (1961, b) the 100 largest corporations in manufacturing, mining and retail (the so called 'giants') were identified in the U.S. economy on the years 1909, 1919, 1929, 1935, 1948 and 1958. For each year, size was valued by the share of firm assets relative to the total of assets of the whole giants' group, as a significant representation of the ability to engage in economic activity, and data were derived from various Reports of U.S. Commissioner of Internal Revenue. A collection of measures was applied, including the simple Markov process, to interpret changes over time within their size distributions and to understand whether even giant enterprises can be subjected to strong competition. Four assets-share ranges have been created by dividing the firm assets domain through a truncated geometric progression and a fifth 'Not-on-List' group has been put alongside them. The first-order, stationary transition matrix, based on those five classes, is generated for every one of the pairs of dates 1909-1919, 1919-1929, 1929-1935, 1935-1948, 1948-1958. Then, the steady state (i.e. the equilibrium distribution) is projected for each transition matrix, trying to go beyond stationarity and grasp the influence of a potential time inhomogeneity over future stochastic previsions. The long-run trends, highlighted for those industrial giants, are the decline in the frequency of identity modifications, as well as of modifications in relative size positions, and the inclination to move closer to one another in relative size. A few months before its printing, this in-depth, cross-sectoral screening had been preceded by a slightly shorter one of the same academics, Collins and Preston (1961, a), focused on relative shares of firm assets of the sole 100 largest food-processing companies in the U.S.A. (from the Reports of the Internal Revenue Commissioner, again), considered over all five-year intervals from 1935 until 1955. A state space of five assets-share ranges plus the 'Not-on-List' category

has been generated by the procedure in the aforesaid article. The first-order time-homogeneous transition matrix has been estimated within each couple of dates 1935-1940, 1940-1945, 1945-1950, 1950-1955: the respective steady-state distributions have been projected to be compared with the actual ones and the values of a five-year index of industrial mobility, similar to the Adelman Index, have been computed to evaluate the turnover. It was concluded that assets distribution, concentration and internal turnover for food manufactures changed modestly.

For Judge and Swanson (1962) the strong postulate of time-homogeneity of the transition matrix is adequate as much as other strong assumptions in many past influential studies involving long-run comparative statistics, so long as the surveyed panel is in a dynamic-equilibrium condition. They tried to portray the paths of annual changes in firms output considering a sample of 83 hog-producing farms in central Illinois over the period 1946-1958 (consisting of 12 observed annual jumps for each farm) for a total amount of 996 observations. Data had been obtained by E. R. Swanson and D. R. Meline from some American institutions, competent in Agriculture, for the drafting of the article published as an AERR-36 by the Department of Agricultural Economics of the University of Illinois on October 1960, and entitled "Hog Supply Response of Farm Bureau, Farm Management Service Co-operators". To qualify for the sample, the farms must have produced hogs in three or more of the 13 years; the number of annually produced litters of hogs is the size proxy. Its Markov equilibrium distribution is calculated, the possible absorption regime is discussed and the Adelman's mobility index, based on mean lifetimes in each state, is estimated. Besides, they first probed in the scientific literature, for an economic variable of a particular industrial sector, the null hypothesis of independence from year t of transition probabilities in year t +1, by the Anderson-Goodman dedicated chi-squared test. A precise chi-squared value, confirming the null and statistically significant at the 99% confidence level, was obtained.

So, since the beginning of the second half of the 20th century, *discrete-time*, *time-homogeneous*, *first-order Markov models* became the stochastic paradigm 'par excellence' to be spent in statistical economics (see Appendix A at page 615 in Zimmermann et Alii (2009) for a schematic, but exhaustive, compendium dealing with implementations in Agriculture). They seemed to

provide both a likely equilibrium configuration for the future (with which to compare the current frequencies distribution) and some plain specifications of the mobility within a system composed by moving individuals, by a transition matrix cross-classifying the states occupied at two different observation periods. They seemed to work properly for incomes (or wealth), as well as for industrial structure and some dynamic phenomena concerning firms. Such a preference had some obvious justifications. Stationary Markov theory is conceptually straightforward and easy to be put in place into algorithms for electronic automatic calculations: over a prefixed epoch the chain is determined by a unique transition matrix, and the progression of time periods after the starting state is represented by accordingly increasing the power of exponentiation on that matrix. It was an advantage compared to the very modest computers' capacity on the middle of the past century. There again, the resorting to more sophisticated theories was being deterred for long by some fear of the danger of circular reasoning, namely the fear that a specific dynamic law, to be identified inherently to the sample, could have been concealed *a priori* among postulates and characteristics of the theory itself (read the 'Concluding Observations' in the accurate Stavins and Stanton (1980)).

In the wake of Judge and Swanson (1962), especially statisticians interested in Agriculture, with an emphasis on crop farms, livestock breeding and dairy factories (and, alongside them, the derived foods productions), were trusting in it. There is a brief list of scientific writings, tackling that subject until the end of the 1970s, at page 5 of Stavins and Stanton (1980). But the third section in Zimmermann et Alii (2009), together with its first table and the table in its first appendix, offers an even more thorough and recent review dedicated to Markov Chains for modelling farm structural change until the end of the 2000s (in subsequent sections of this article, not-Markov econometric models and the new multiagent-based procedures are reviewed). The two afore-mentioned tables are respectively dedicated to the time-inhomogeneous chain and to the time-homogeneous one. In particular, there have been arranged inside them for each scientific publication, besides the region, the Agri-food subsector and the economic size proxy, also the type of data (micro or macro) and the analysed epoch, the number of classes constituting the backing space and the recourse to some input-output repository, the inferential method to estimate the parameters as well as potential exogenous variables driving the population dynamics and,

finally, the performance gauged by the coefficient of determination R^2. Another more synthetic review of this kind of literature is in the first table at page 3 in Piet (2008), where the coefficient of determination, the inferential technique and exogenous variables are not indicated while historical lapse is substituted by the transition frequency and the type of data is distinguished into individual (micro) or aggregated (macro).

In Bostwick (1962) it is expressed the persuasion that wheat yields in the Montana state of U.S.A. can be fitted by the simpler Markov process, thanks to its ability to include traditional autocorrelations between yields in the dryland cropping systems. He gave quite attention to the mean first-passage times and he also attempted to illustrate some examples of application to farm financial management, namely to the strategy of cash carryover, to the fertilizer decision making and to the allocation of surpluses.

In Padberg (1962) an effort was made to understand developments of the wholesale fluid milk industry in California respect to the sales volume. Data had been obtained from the California Bureau of Milk Stabilization for the months of January on 1950, 1955 and 1960, that delimit the only two five-year periods 1950-55 (when the panel consisted of 241 companies) and 1955-60 (when the companies reduced to 176). There are two rough approximations: the brevity of the sequence of observations about the evolution of the chain, should be compensated by the pluriannual extent of the periods; the use of a not-fixed panel of plants had been believed enough to incorporate mergers and acquisitions inside the dataset and departures from it. Factories had been decreasingly ordered respect to their output and the adopted size categories were aggregate market shares, preferred to absolute-values classes for sales or assets because variations on shares could better represent movements within the size distribution and changes in market structure. The last fourth class of the lowest shareholders was exploited as a storeroom capable of better handling the entering, the exiting and the merging agents. At the end of the work, a likelihood ratio test (already published in 1954 by T. W. Anderson in "Probability models for analyzing time changes in attitudes") was executed about the hypothesis that forces affecting firms' growth had been identical on 1950-55 and on 1955-60, ergo transition probabilities were the same over both periods (*stationarity* of the chain). Surprisingly, such null of stationarity on the decade 1950-1960 is rejected, anticipating the refusal of time independence

which would have resulted, from the half of the 1970s onwards, in many screenings respect to different variables by different researchers. Since Padberg himself had observed that the data-estimated likelihood ratio may increase with the number of states, he ascribed the rejection to the small number of category (only four) composing the state space. He concluded the *first-order Markov stationarity*, along with its equilibrium distribution, *in the long run approximates tendencies* of manufactures' dynamics, *in case the environment to remain unaltered*, *without exactly foretelling* it.

The Markov process is the ultimate probabilistic technique to disclose the structure of markets in the researches of Williams and Alexander about fluid milk markets in Louisiana. Milk handlers and milk producers were considered separately in Louisiana, which was in its turn divided into the three areas of New Orleans, the Northern region and the Central region. The volume of sales in pounds was selected to represent handlers' size, the pounds of output to represent producers' size. Sources of data were many and heterogeneous: State Market Administrators, the United States Census and milk producer associations. We take the opportunity to express the opinion that *lumping together a too heterogeneous set of partial data sources* may be a problem. It *might originate a fictitious sub-sample of firms* for the analysis, *which has nothing to do with any sub-sample from the real population of firms to be assayed and, as a consequence, some outcomes might be falsified*. The extremely wide and meticulous survey by Williams (1963) blended systematically together all the peculiarities of the most of previous references. In particular, structural modifications from 1951 to 1962 were characterized in relation to average number and average size of firms; size distributions and growth patterns in the last decade were estimated along with future size distributions of firms in case of dynamic equilibrium; mean lifetimes and the Adelman's index of firm mobility was derived and compared to a perfectly mobile industry *à la Prais*. Albeit the inspected lapse went from 1951 to 1962, transition probabilities and mean lifetime were estimated only from 1956 (or 1958) to 1962; the Authors have also ventured into some projections for 1972. More than once, some chi-squared tests were executed, confirming the dependence of size distributions at time $t + 1$ on the ones at time t and excluding differences between transition matrices extracted from different milk markets. The rejection by Padberg's chi-squared test was hastily cited as an

error which had been amended, but we have not tracked down anything published about such revision. However, the execution of cogent chi-squared tests directly conceived about Markov Chains, either for the time-homogeneity or for the order, was unhelpfully avoided. Mirroring the guidelines in Williams (1963), the little more synthetic exposition by Alexander (1965) paid new attention to the connection between growth patterns and chances of survival, as well as chances of variations in sales' volume for milk producers.

Krenz (1964) spent this model to try foreseeing in the future (or '*projecting*', as it will be equivalently said in the successive scientific literature) the number of farms in North Dakota on 1975 and 2000, starting from different points in the past. He was the first estimating a stationary Markov model from macro-data (or 'census data', or even 'general aggregated data') by an intuitive procedure, which necessitates of some further assumptions to be adjoined beside the Markov ones and is parallel respect to the Anderson-Goodman classic mathematical inference requiring micro-data (or 'raw data', or even 'granular data', which are qualified individual by individual and time after time, further severable into time-series data, cross-sectional data and panel, or longitudinal, data). Data are drawn from the Quinquennial U.S. Census of Agriculture from 1935 to 1959 where farms had been classified by their acreage in conformity with the U.S. Census Bureau definitions. The Author himself highlighted some difficulties implicit in his approach, like the too much extensive and demanding assortment of suppositions sustaining it and the impracticality of any reliability check-up on the inferred transition probabilities.

In the second half of the 1960s some researches (among the pioneering ones there are Lee, Judge and Takayama (1965) as well as Lee, Judge and Zellner (1968)) demonstrated it is scientifically possible, via econometric estimation, to build robust Markov models from cross-sectional data only. So, Markov inferential methods and applications resting on macro-data (which are more common, less expensive and less difficult to be elaborated despite they must be pondered with caution) have been getting more and more attention until today, especially in Agri-food economics. In the following chapter, we shall mention some other publications of this kind (Chavas and Magand (1988); Disney et Alii (1988); Piet (2008)), but they are out of the scope of the prosecution of our work. On the contrary, in the most of our references, both the Chains and the χ^2-type tests are based on elaborations of panel micro-data (sample

longitudinal data) or on sets of empirical points from time series. The empirical analysis itself in the second part of our thesis (Chapters 6 and 7) involves longitudinal micro-data relative to almost complete, fixed sub-populations (so the panel is balanced) of some Italian manufactures (see Chapter 6). Another vast strand of research, which sprang from Krenz's commitment and is yet very active, obviously deals with Markov projections not only of the number of farms, for a variety of size proxies, in some European state as well as in American or Asian regions, but also the number of corporations in various economic sectors: we cannot account for it, except for Hallberg (1969) and few others in the next chapter.

Moving to examinations about manufactures, in Horowitz and Horowitz (1968) an innovative perspective is proposed. *Without executing any check-up*, consumer purchases among firms in the brewing industry is assumed to follow a first order, stationary Markov process and, using an alternative, quadratic-programming technique by Theil and Rey, they estimated the transition probability matrix. Then, recalling the idea of some communications scientists (e.g., E. Shannon), they suggested establishing an analogy between the concept from information theory of entropy of a source, on the one hand, and the phenomenon of the industrial competition, on the other; businesses' concentration is, in its turn, strongly positively correlated to competitiveness inside the sector. Concentration in the U.S. brewing industry between 1944 and 1964 was analysed with respect to the entropy measure which had become a direct proxy for the degree of competition. Data had been taken from various issues of the Brewing Industry Survey and had been distributed upon a state space, consisting of six size-categories of its most representative U.S. manufactures. A rather unique picture of the trends and degree of concentration and competition in the industry evolves from this analysis. Further, they derived a transition probability matrix, involving most of the leading brewers and reflecting shifts in market shares within the industry, to analyse trends in the industry. Authors were aware that *results were strongly influenced by the selected groupings and* that transition probabilities were estimated including *the effects of mergers and acquisitions* (the number of firms declined from 374 in 1944 to 129 in 1964), with no possibility of gauging the effect over the transition matrix if mergers had not taken place. *The latter is a frequently*

recurring flaw in the statistical estimation of transition probabilities, in the attempt to render the sectoral dynamics.

Chapter 2
Historical synopsis of Markov Chains in Economics: a second age of novelties

2.1 A new critical attitude onto Markov time-homogeneity in the 1970s and 1980s

The turn of the 1960s inaugurated a second age of applications, when it began to be questioned the efficiency of first-order stationary Markov chains, at least in the case of firm-size proxies, and to seek out to refine the framework. These efforts were done, to quote just a few examples, by inserting a random mechanism to reproduce movements of some agents between the active sample and the passive repository of the state space; or by introducing into the chain some kind of dependency on the history of the system (a not-stationary transition matrix built up on the dynamics of some exogenous forces, an order of the chain equal to two or higher); or by enhancing the theoretical layout (e.g., mixtures of diverse chains, instead of a single one, like the mover-stayer model; integration of a logit model into the estimation of transitions probabilities; chains which are continuous, instead of being discrete, in time). Some Authors began to compare among one another diverse Markov models or different scenarios for the same model. Remember the first table at pages 605-606 in Zimmermann et Alii (2009) for an overview on the most important, not-stationary Markov-type contributions regarding Agri-food enterprises. Chi-squared tests and maximum likelihood ratio criterion began to be performed with increasing frequency (and, sometimes, they were repeated for multiple scenarios), onto increasingly ample databases, in research plans intended for publication. The new attitude was sustained by gradual but relentless innovations in computing technologies.

Among the first investigations, taking publicly a stand, there was Hallberg (1969). The Author asserted to have found, by the maximum ratio criterion for time dependence in Anderson and Goodman (1957), to be inappropriate the assumption of the transition probabilities remaining constant over time for the

annual sales volume of 884 plants which had manufactured frozen milk products from 1943 to 1963 in Pennsylvania. First-order stationarity led to erroneous predictions respect to 1964 and 1965 real data. Hence, a method involving multiple regression techniques was elaborated to replace constant-in-time, first-order transition probabilities with functions of a selection of time-varying factors (whose values were deflated, in most of cases). Such factors corresponded to some connotations of both the industry (indices of workers' hourly earnings in U.S. food manufactures and of retail prices for U.S. dairy products, public funds paid to Pennsylvania farmers for all milk produced) and of the entire state market (population in Pennsylvania and state per-capita income). New time-varying transition probabilities were estimated by means of the least-squares strategy; better previsions, than by the time-homogeneous migration mechanism, were achieved. Data had been registered by the Crop Reporting Service of the Pennsylvania Department of Agriculture and the 884 had been in operation for at least one year during the double decade but not all 884 in the same year. Thus, a last fifth state, consisting of the plants in the sample producing no output during a given year and akin to a simple entry-exit repository, was placed next to the active four classes of positive outputs (in gallons), accounting in this manner for the dynamics intrinsic in the industry.

Anyhow, starting from the observation that the real evolution of an industry (and consequently the development of credible predictive models) is made complex by repeated, previously neglected entrance and departure events of some firms relative to the market, in Duncan and Lin (1972) the Authors proposed an extension of the 1st-order, stationary Markov chain through stochastic entries and exits. It was applied to the ratio of bank's loans disbursed to farms to the net total loans for the Ninth Federal Reserve District banks from 1954 to 1969 (data had been taken from December Call Reports submitted by each member bank of the Ninth District). The maximal size reached by the sample is 528 banks and predictions were made from 1970 to 1975. Five loan ratio strata and a unique *zero*th absorbing state for all outputs without opportunity of re-entry, next to them, were determined; many forms of stochastic entry mechanisms into any one of the active strata are contemplated (renewing binomial, negative binomial, Poisson), even if great care is given above all to the multinomial-entries process. Since the total likelihood function of the whole system, comprising new entrants at every period and the

absorbing state for the leaving farms, can be factorized, the sufficient statistics are the counts for all jumps (as for the case of a fixed sample) so the maximum likelihood estimates and likelihood ratio tests are formally the usual ones resulting in Anderson and Goodman (1957) for multinomial parameters. Regarding likelihood ratio tests, the Authors wrote that such controls had fully confirmed the joint hypothesis of a 1st-order, stationary Markov system against the alternative of 2nd-order and time-dependent transition probabilities. Pearson's goodness-of-fit tests too, comparing predicted frequencies distributions with the observed ones in 1968 and in 1969, had been in favour of the 1st-order stationary Markov model extended by stochastic entries for new firms together with the inactive container for departures from the industry.

Nevertheless, more than twenty years after Champernowne's final 1st-order Markov model combined with the law of proportionate effects for incomes, Shorrocks (1976), relying on evidences for the observed transition matrices from a British fixed sample of 800 coetaneous male employees, whose annual incomes are known for the years 1963, 1966 and 1970 (not for the intermediate ones), recast doubts on whether the mobility is a time-independent first-order Markov process, as a rule, in economics for income variables. After some preliminary considerations, both theoretical and empirical, a generalization was proposed by reinterpreting Champernowne's hypotheses (law of proportionate effects included), which leaded to a second-order stationary Markov chain. Maximum-likelihood estimates of the transition probabilities were computed in case of mobility restricted to a single stage movement in each possible direction; some performed likelihood ratio tests showed the second order to be an improvement (curiously, Shorrocks did not cited the work by Anderson and Goodman as a reference about these tests). Though, he renounced inspecting whether such deficiency of the discrete-time, first-order, stationary Markov process is intrinsic to the model or rather it was caused by some great-impact events occurred during the period from 1963 until 1970.

By the way, it is well-known that any 2nd-order stationary Markov chain, built up on a set of m elementary states, can be represented as a more complicated 1st-order chain over a set of m^2 composite states, everyone made up by a couple of the previous elementary states, underlying an m^2-square transition matrix where a large number of conditional probabilities are necessarily equal to '0'

(see Anderson and Goodman (1957)). But such an approach has been considered less attractive, so far, in socio-economic sciences.
Instead of taking for granted any exegetical supremacy of the simple stationary chain, maybe upgrading it by a stochastic entry-exit process for companies which are not permanent inside the core of the sample, in Stavins and Stanton (1980) a host of Markov first-order schemes has been reviewed, also via statistical comparisons respect to the same data. All these variants were obtained by matching three different type of data (micro data, macro data and aggregated ones from time series) with the two variants of the transition matrix (the time-homogeneous matrix and the time-dependent one), as well as with some procedures to project a not-stationary chain in the future. From this host of models, three distinct, first-order frameworks were implemented on the set of data at disposal from the New York State dairy sector: the stationary chain based both on micro data and on macro data, whose parameters have been respectively estimated in compliance to the Anderson-Goodman inference and the Krenz rules, then a micro-data multinomial logit model resting on transition probabilities, varying with the time, and on the milk-feed price ratio as an explanatory variable. In-depth comparative studies, like this one, are seldom conducted in Economics and it has inspired some traits of the empirical analysis in the second part of our thesis. Data pertained to twenty counties in the State of New York where milk was sold under the New York-New Jersey Milk Marketing Order, from January 1968 until December 1977. Throughout the decade, 14,272 farms had sold milk at some years in the area, from which a subsample of 1,012 permanently producing plants was extracted. Such permanent plants were sorted into 10 categories defined by milk pounds sold per month per farm, whose breadth is 20,000 pounds for everyone: the lowest, zero-production category accounting for entries and exits, eight intermediate finite classes and the last open-ended one. In order to compare to each other the three above-mentioned, first-order models and to figure out which of them fitted as best as possible the “real world” developments, some simulations were made for the year 1977 using data of the subperiod 1968-1974. The micro-data multinomial logit model produced the most precise simulations of the 1977 actual size distribution; the second better performance was by the stationary, micro-data, Markov model; the macro-data, Markov approach by Krenz was the worst. Concurrently, the null hypothesis of chain’s time-homogeneity on

the same subperiod 1968-1974 was tested *à la* Anderson-Goodman and rejected, providing a strong confirmation of the time-dependence of transition probabilities for dairy productions, as already prefigured in Padberg (1962) for the 1950s in California, in parallel with what had been verified for British incomes during the 1960s in Shorrocks (1976). Nevertheless, the micro-data time-homogeneous chain sounded to perform better than the stationary Markov chain *à la* Krenz anchored in macro-data. Finally, parameters of the multinomial logit model have been re-estimated upon the overall 1968-1977 dataset in order to yield annual predictions from 1978 to 1985 concerning New York State dairy farms, in the context of four distinct scenarios involving the milk-feed price ratio (constant, constantly increasing, constantly decreasing, historically fluctuating).

Parenthetically, we point out to have found no scientific publications verifying later the fidelity to the real facts of Markov projections which had been made for the future by any one of our heterogeneous references. This is a weakness of economics research and it is one of the reasons why we have opted for systematically testing the credibility of the first-order Markov stationarity in the case of some Italian industries at a national scale.

Remaining inside Agri-food industry and retrieving Hallberg's general scheme, but coming back to the cereal sub-sector, a sketch of comparison, without any formal check, has been also developed in Mellor (1984) between the first-order stationary chain and an upgrade of it. The former has been estimated both by the Ordinary Least Squares (OLS) technique and by the Restricted Least Squares one (RLS) on a linear regression involving the unconditional frequencies of occupying in the arrival class and the departures ones, respectively at the dates t and $t-1$, besides the stationary transition probabilities and a vector term of disturbances. The upgrade embodies linearly the influence of some exogenous variables, carrying with them time-dependency into the transition matrix. We must outline that the time-varying jump probabilities have not been explicitly calculated and the determination of the linear, not-stationary upgraded model remains, in a sense, unfinished. Moreover, some negative values of the OLS-estimated stationary jump probabilities, along with some negative values for the RLS-estimated parameters of the time-inhomogeneous upgrade, have left us perplexed. However, annual rotation of three traditional cereal crops (wheat, barley and oats corresponding to three

pairwise-disjoint states) over a unit of arable land in Great Britain was studied according to aggregate time-series data (namely, the number of units in each cereal category at each date, or a temporal sequence of census data), from 1866 to 1978 and, separately, on the sub-period 1945-1978. There exist a fourth, 'not-cereal crop' state in order to close the market, so this survey is one of our few references grounded upon a qualitative state space and where 'statistical individuals' are unit areas instead of economic agents. For every cereal's category and every observation date, an adaptive expectations mechanism (with coefficient 1) has been applied to profitability-per-hectare, which is defined as the share, relative to the mean price, of a linear function of returns-per-hectare (i.e. a combination of price-with-yield products). Not-stationary transition probabilities have been defined as linear functions of relative profitability for every cereal class, at the previous date. Three systems of linear equations, owing to three different levels of constraints, have been *solved via an unknown software*. In conclusion, the time-independent transition matrix sounds to be disfavoured, because of "a significant response by the transition probabilities to exogenous variables (page 215)". Exogenous variables have to be taken into great consideration to lend a better forecasting power to the Markov model, albeit the Author himself acknowledged that such methodology, embodying the influence of time into the stochastic chain, should be improved.

Ethridge with his co-authors has maintained the comparative plot between the two first-order chains to analyse, from 1967 to 1979, changes in number and size of 376 cotton gin firms, spread throughout a 23-county area in Texas High Plains in the article Ethridge, Roy and Myers (1985). Interestingly, he has confessed some problems arose during the estimation of the time-inhomogeneous transition matrix from data of the U.S. Agriculture and Commerce Departments, owing to the very limited sample's size compared to the number of categories of Markov space. Productivity, intended as the number of bales-per-hour, has been used as indicator of cotton-gins' size. Ethridge is our sole reference constructing 12 composite classes, exhaustive and mutually exclusive, by pairing five firm-size groups both with the group of active ginning factories and with the group of the not-operating gins, then by adding the two classes of new entrants and dead gins since 1967. He was also one of the first academics who noticed and published the stylised fact of the

clustering around the main diagonal inside the time-homogeneous transition matrix of an economic size-proxies between two consecutive years. It consists in the overall tendency for most of the active factories to remain in their original status and it is interpreted as a kind of inertia, intrinsic to the size dynamics of any real economic agent. In order to estimate the not-stationary chain by the least-squares procedure, every element of the twelve annual transition matrices was considered as the dependent variable for a specific regression equation, provided that at least eight observations exist for the fixed jump. As predictor forces, conditioning the time-varying probabilities, six proxies have been respectively adopted for changes in labour costs and in energy costs; for percentages in seasonal plant capacity and in county production's change; for both periodic and gradual technological advances. Authors were not able to conduct any chi-squared test on stationarity, though it would have been helpful, because there were no observations for many yearly jumps, owing to the great number of classes along with the quite scarce number of statistical units. *In general, an inadequate number of observations for some jumps is a serious trouble pertaining statistical inference about Markov chains, especially for the time-dependent one*. Some simulations, based on the 1979 distribution, have been executed quinquennially from 1984 until 1999 about future distributions using both the first order counterparts, under a variety of evolutionary regimes for labour cost, energy cost and periodic technological progression. At the end, Authors declare their preference for the not-stationary Markov chain despite it requires a very detailed ensemble of data to yield reliable and complete projections.

Really Hallberg's, Mellor's and Ethridge's viewpoints had been anticipated by intuitions in Telser (1962) about consumer behaviour, where it had been admitted that the choice of a brand is correlated to the advertising disbursals of the producing company and this relationship influences the probability of repeat or switch purchase. Other important exogenous determinants varying with the time (like prices) may impact, pace by pace, on jumps.

The layout of the latter four Authors has been borrowed, at the end of the 1980s, by Chavas and Magand (1988), Disney and Duffy and Hardy (1988) (or Disney et Alii (1988)), matching the endeavour of the Markov Chains' estimation onto macro-data (let us remind Krenz (1964)) with the task of a time-inhomogeneous formulation for the first-order chain, in virtue of

exogenous determinants' intervention. In both of last econometric prototypes, transition probabilities $p_{ij}(t)$ are expected to evolve also according to an ensemble of external factors $\{x_t\}$, presumed to be the ones more influencing firm production (and the entry-exit process, too), in virtue of some function f : $p_{ij}(t) = f\ (i,\ j,\{x_t\})$, and where the most impacting are costs, efficiency and prices (or ratios between apparently uncorrelated prices. The future distribution $s(t+1)$ depends on the current distribution $s(t)$ and the transition matrix, save for a vector $\varepsilon(t+1)$ of error term: $s_i(t+1) = \sum_j p_{ij}(t)\ s_j(t) + \varepsilon_i(t+1)$. In practice, it is easy to solve the derived systems of equations when $f(\cdot)$ is linear respect to the external factors, although it is not conceptually necessary. It remains a sure fact that no credibility checks are possible pertaining census quantitative information. In Chavas and Magand (1988) the focus is, for the umpteenth time in the economic literature, on U.S. dairy production and data have been extracted from nationwide, state-by-state surveys of the U.S. Department of Agriculture, for the period 1977-1984. Milk production, in four interstate macro-regions, has been assessed by the number of cows per farm and classified by four herd-size categories. The model is grounded on a regressive linear relation for the number of farms in the fixed i-th category, in the State k, at the date t respect to the number of farms, in the same State k and in all four categories, at date $t-1$, and respect to the number of net entries. The adopted estimation procedure is the Seemingly Unrelated Nonlinear Regression (SUR) with zero-mean error terms, whose covariances are zero for pairs of different states and different dates. Explanatory determinants are sunk costs, scale efficiency and market prices. Disney et Alii (1988), instead, by the Markov first-order not-stationarity have provided an insight on indirect impact of grain pricing policies on pork farming. In our bibliography, it is one of the very few investigations undertaking how and how much two distinct commodity sectors could bear upon each other. Information were acquired from pooled U.S. Census data on 1969, 1974, 1978, and 1982 across five Census interstate divisions (accounting for 96% of U.S. pork production in 1982). The state space is made up of four market categories, according to the number of hogs sold per year, plus a possible, fifth exit category. The transition matrix, linearly regressed at each date onto the hog-corn price ratio, has been estimated by linear programming combined with Minimization of Absolute Deviations

(MAD). Such fundamental econometric frame has been split into four different models by admitting both proportional disappearance among the four selling classes and not-proportional exit events, when the hog-corn price ratio is conditioning the jumps' dynamics, as well as in the case of no effect from that ratio. Finally, the sensitivity of the pork industry to the hog-corn price ratio has been recognised to be relevant (despite of pork production and corn crops seeming to be weakly correlated) and simulations of projections, to the year 2000 under diverse price regimes, have been executed.

Contemporaneously, during the second half of the 1980s, some academics went on trusting in stationary Markov chains, but in the case of more sophisticated contexts, at an upper level of analysis. Neftci (1984) and Nelson, Braden, Roh (1989) gave an instance of *application of the Markov first-order stationarity to unfold* not the collective dynamics of companies respect to an economic variable, but rather *any possible asymmetry, relative to a binary proxy, hidden inside aggregate time series*. The same stochastic plot exploited by Neftci (1984), to see whether an asymmetric behaviour is exhibited by unemployment rate over the phases of a business cycle (we shall not dwell further on this publication), can be tracked down in Nelson, Braden, Roh (1989) for tackling the issue of agricultural assets' fixity. In the latter, irreversibility of investments in agricultural capital stock (namely, the fact that it is more difficult to dispose of the assets for a particular production than to add new capital to their overall current stock) is explored within the Markov framework by directly and systematically testing the *null hypothesis that asset-selling operations persist longer than operations to acquire new productive stocks* (*asymmetric asset adjustment*). The subtending space is made up of two states, the investment represented by a positive or null increment $X(t)-X(t-1) \geq 0$ together with the disinvestment represented by a negative increment $X(t')-X(t'-1) < 0$, where $X(t)$ is the stationary transform of a raw capital-stock series. Then, an auxiliary binary process $I(t)$ is defined as equal to 1 in case of investment and equal to 0 otherwise. For the Authors, when investment irreversibility befalls, the process $I(t)$ is expected to stay in the 0-state longer than in the 1-state; so, asset fixity is investigated testing whether the spurious drift from 0 to 0, whose conditional probability is λ_{00}, is more likely than the spurious drift from 1 to 1, whose conditional probability is λ_{11}. Probabilities of jumps have been estimated by the Maximum Likelihood principle; probability

asymmetry has been tested by Lagrange Multipliers; Wald test has been applied to the null hypothesis. Data were obtained from the U.S. Department of Agriculture and Department of Commerce and consist of annual time series for the net capital stock of nine asset categories (including real estate and non-residential buildings, motor vehicles and metalworking machinery, trucks, buses, tractors) and for the gross capital stock of seven of those nine categories. Starting years are 1913, 1925 or 1947 and ending years are 1984, 1985 or 1986, depending upon the asset category; the series are 24 in total. A comprehensive table listing estimated jumps' probabilities is at page 975; another comprehensive table for the 24 test-statistics' values and associated *p*-values is at page 976. *Such a broad range of test values, spread over a variety of variables, is almost an exception in the economic literature and it inspired the tables of reliability assessments in Chapter 7 of this thesis. However, we have not verified any kind of asymmetry, because of the brevity of time series in our datasets, but rather some Markov attributes between each other, leveraging the sampling extent of the archive at our disposal.* *P*-values of 0.10 or less have been chosen to represent strong evidence of asymmetry; *p*-values greater than 0.10 but less than (or equal to) 0.20 have been chosen to indicate weak evidence of asymmetry. Lastly, through the Markov Chain approach, only twelve of the twenty-four agricultural time series have exhibited a statistically significant asymmetry; moreover, albeit the following statements are not statistically incontrovertible at the sight of the *p*-values, asymmetry in agriculture seems to affect slightly more the net capital stock, to be more pronounced for specialised asset categories, to characterized the period after World War II more than other longer ages during the 20th century.

2.2 From the 1990s to present days, emphasising microdata and time series

Twenty years after Shorrocks's forerunner criticism on the ability of the stationary, 1st-order, Markov process to describe income dynamics in Britain, doubts were borne out about it by Schluter (1997), on the non-stationarity of German income mobility according to the data of the German Socio-Economic Panel from 1983 to 1989. The Panel contained two types of incomes: granular data about annual pre-taxes, pre-benefit income and estimations, derived from

simulations, on annual household post-taxes, post-benefit income. Only persons having a complete income record for all the years were selected for the analysis, resulting 9022 observations. He applied to the panel various Markov models, comprised the stationary mover-stayer model and the sketch of its not-stationary homologous, and asserted to have controlled by the maximum likelihood criterion a set of pairs (null hypothesis-alternative one) on the time-dependence of the transition matrix and the order of the chain, following the lesson of Anderson and Goodman or their spirit. From such tests it is clear the not-stationarity is far more likely than time-independence as well as the second order is more probable than the first one for a chain. Inter alia, after observing that the distribution of German incomes remained almost unchanged during the 1983-1989 period, a careful study of the non-stationary transition matrices, along with the values of the trace mobility index based on them, reveals substantial movements beneath the 'almost inactive surface' of the distributional shape.

Bickenbach and Bode (2003) claimed even more decisively against the basic Markov hypothesis, in a 'regional' geographical scale and in the context of economic convergence analysis based on the Markov limiting distribution. They have carried out chi-square and maximum likelihood ratio tests of the Markov property, of spatial independence, and homogeneity over time and space for first-order transition probabilities (inspired by the paper by Anderson and Goodman, but where some corrections were brought about when it befalls some observed counts, parameterised at the denominator inside the statistics, are null) focusing on relative regional per capita incomes in the United States. The result, pertaining 3408 aggregated observations derived from the available time series, has been that "the evolution of [relative, regional,] per capita income distribution across the forty-eight contiguous U.S.A. states from 1929 to 2000 clearly does not follow a common [discrete-time,] first-order, stationary Markov process". To avoid any 'poor statistics situation' during the data analysis, periods longer than one year are adopted even if the time in the data is annually scanned. It is interesting to note that the Authors' analysis attributes the failure, beyond the restrictive nature of the Markov assumptions, to various kind of reasons, due to complex interactions among the states inside the U.S. federation as well as between U.S.A. and the international scenario. Firstly, two breaks occurred in the historical period under study: the major in

the aftermath of World War II, that significantly conditioned the evolution of the income distribution; a later minor one in the late 1990s. Then, a different development is exhibited by certain groups of states compared to other ones; finally, states surrounded by poor neighbours show a different development than states with rich neighbours.

In the previous literature on Markov Chains, the transition matrix is almost always built by discretizing the population of firms (or households) into a limited number of categories according to some peculiar criterion. Interestingly, in Piet (2008) some drawbacks of Discrete Markov Chains are underlined. The small number of intervals (very often from 5 to 9 or, very rarely, to a number slightly bigger than 10), discretizing the variable domain, makes the transition matrix either to be simply diagonal or to exhibit a concentration of most of the probability mass around the primary diagonal, so important information is lost about the fine structure of individual drifts, as the time passes. In addition, discretization may lead to the counterfeit conclusion that the population distribution can become bimodal, what is an artefact recurring when working with histograms, due to the definition of the size-interval bounds. Just to avoid these problems, Piet (2008) has proposed a *stationary but continuous first-order Markov Chain* as a more informative framework, enabling to derive more in-depth predictions about the population's distribution. The goal is to describe the population's dynamics of professional farms as represented by aggregate data of the panel for France from the FADN (Farm Accounting Data Network), relatively to the Utilised Agricultural Area (UAA) as a size proxy. Holding the Gibrat's Law of Proportionate Effects (whose formulation is $h_i(t+\tau) = (1+\partial h)\cdot h_i(t)$, where $h_i(t)$ is the size of the i-th company at time t and ∂h is the growth rate over the period τ) to be a good approximation confirmed by French data, it has been supposed that the empirical cumulated distribution of the UAA from 1980 to 2005 could be fitted by a two-parameter lognormal probability distribution (in other words, that the number of farms $n_h(t)$, exhibiting a fixed size h at a fixed date t, followed a two-parameter lognormal density function). Besides, the size probability density $P(h_i(t+\tau))$, upon which it has been hinged the convolution defining the evolution of the size distribution of farms, has been supposed to have the shape $p(\partial h;\tau)$, depending only upon the growth rate ∂h and upon the elapsed time τ.

Such $p(\partial h;\tau)$ must replace, in the Continuous Markov Chain, the conditional transition probabilities of the Discrete Chain. The last plausible assumption has been made at a macroeconomic scale: independence of the size change of any farm in a region, between two separate dates, from the size variations of other farms in other regions. In order to be able to turn back to any discrete transition matrix, which is specified by a ratio between integrals involving $p(\partial h;\tau)$, $n_h(t)$ and the bounds of the transition classes, for the same growth-rate density $p(\partial h;\tau)$ another three-parameter lognormal distribution has been adopted. Estimation of all parameters inside the two posited lognormal distributions have been obtained through the nonlinear least-squares procedure in the Stata 10.0 software. Some projections have been executed towards the year 2015, taking advantage of the ten-year probability density $p(\partial h;\tau = 10)$. The not-stationary development of any Continuous Markov Theory for socio-economic systems, along with the integration of statistical units entering a market or leaving it, and the mathematical foundation of suitable tests assessing the trustworthiness of Continuous Chains respect to granular data and aggregated data, are open problems.

As a novel mode of application of the simple Markov chain to capture the impact of political actions, promoted by an international authority, on stakeholders' decision-making processes, conditioning in their turn industrial evolution, we like to cite Bertoni, Aletti et Alii (2018). They suggestively are searching for asymmetries, raised from a possible turning point, within the stationary Markov framework of the first order via comparisons between its translocation probabilities (in a diverse way from Neftci (1984) and Nelson, Braden, Roh (1989)), offering some food for thought. It is also appealing for the resorting to a weighted version of the chi-squared test for contingency tables filled with interclass-migration frequencies, and to the Gini-Simpson Index of Heterogeneity, bringing back to mind the components of some mobility indices reckoned by "*MarkovInfer_MobInd.m*". The paper is an ex-post study (i.e. a gauge, not-predictive study), driven by farmers' detected behaviour, of the potential influence of the new 'Greening' program, within the European CAP funding, on utilisation of crop areas. It has been tested via Markov inference whether farmland-use dynamics has changed after the current CAP version has come into effect in 2015. The 'Greening' is a flow of direct payments decreed by the European Union in its last Common

Agricultural Policy (CAP 2015): namely, it is a program governed by rigid rules and assigning 30% of all-round CAP subventions to farms for the improvement of the use of natural resources. A dataset of about two million of arable crops parcels, measured in hectares (ha), over the lapse 2011-2016 (split into the sub-periods 2011-2014, the one preceding the Greening, and 2014-2016, after the Greening) in Lombardy, one of the most powerful Italian regions in a plurality of economic sectors, has been extracted from SISCO, the information system managing farm demands for CAP payments. Only the 638,952 parcels recorded in all the observation years have been considered, extended over 743,072 ha and spread over 23 different categories of crop typologies; it represents almost the universe (always more than 90%) of Utilised Agricultural Area (UAA) in Lombardy. Inside the sample, possible discontinuities, referring to the sub-period 2011-2014 before the Greening and to the subperiod 2014-2016 after, have been sought in parcels' migration phenomena among the 23 crop classes, quantified by the transition probabilities from any of those categories to another one, estimated for each pair of consecutive years. Choices about data by Bertoni, Aletti and co-Authors share two common points with the present doctoral dissertation, besides the systematic redo of χ^2-type assessments for many different contexts commented in Chapter 7. Indeed, our elaborations rest on the extraction of all permanent firms (it is underlined in Chapters 2 and 3) from original microdata electronic sheets of every Italian industry we have examined for diverse variables. In addition, from the official archive AIDA, commissioned by the Italian Union of Chambers of Commerce, we succeeded to export complete sectoral samples, at a national scale, relative to very many combinations of selection criteria: they almost coincide with the real population of a certain Italian industry for those criteria (Chapter 6). As for the check of statistically significant discontinuities, due to the Greening, in any transition probability $p_{ij}(t)$ from the cultivation of type i towards the cultivation of type j between the years t and $t+1$, some preliminary inspections have been needed. The null hypothesis of stationarity of the chain on the sub-period 2011-2014, along with the null of equality of successive transition matrices, subtended by consecutive pairs of consecutive couples of years, have been tested by the Maximum Likelihood Ratio Criteria in Anderson and Goodman (1957), *initially adopting single hectares as* "*panel's individuals*" (*instead of single farms, or single firms in*

general, as we have done in the second part of the thesis, as well as the most of our bibliographical references did!): *p*-values always lower than 0.0001 have led to rejecting the previous two null hypotheses. Authors have justified such rejections, when individuals are parcels or hectares, by the very high degree of geographical correlation among hectares (belonging to the same farm or to different farms near to one another) in the very intensive agricultural sector of Lombardy; by those changes which necessarily recur in land use (e.g., crop rotation); by the great sensitivity of the χ^2 tests to small deviations within large samples (remember that the panel size is 743,072 ha: it is an issue examined in Bergh (2015)). Stationarity of the chain is an essential property to deliver by comparisons the potential discontinuities, hence the definition of "statistical unit" has been reshaped through the concept of "aggregate of hectares maximizing the likelihood function" under the assumption of time-homogeneity of transition probabilities during 2011-2014, in order to centre around the "new agents" an acceptable stationary Markov model. *Such a reshaping, which implies time-homogeneity, does not have to be interpreted as a 'forgery', rather as an overall transformation into a stationary perspective enabling some kinds of asymmetries to unfold themselves. It is a very intriguing topic for further reflections in the scope of statistical economics*! After the redefinition, they have taken advantage of a weighted χ^2 test for the contingency tables of transition frequencies of every starting cultivation class *i* into the others, to filter out physiological land-use inhomogeneities. For all ante-Greening traditional monocultures, new weighted *p*-values for annual transitions of the biennium 2014-2016 are not only manifestly lower than new weighted *p*-values for annual transitions on the sub-period 2011-2014, but also manifestly lower than the prefixed threshold of 0.1 associated to discontinuities in crop reallocations owing to the Greening program. Moreover, from 2014 onwards for the same traditional monocultures (chiefly for maize groups), annual values of the Gini-Simpson Index tend to increase, upholding displacement phenomena to other cultivations (like soybean). It has been corroborated, at a general level at least in Lombardy, the novel Greening policy rules, entered in force concurrently with the 2015 EU CAP, had an impact on agri-food production, favouring cultural diversification.

2.3 Suggestions to understand the puzzle about Markov Chains in Economics

Evidently, the scientific literature displays conflicting judgements toward the hypothesis that the first-order stationary chain is able to describe patterns of the economic variables, checked up in disparate backgrounds: it is natural to wonder what the causes of such a contrast are and whether it is possible to glimpse some regularities. Here we put forth a list of remarks that are as many advices for further inspections about implementations of Markov Theory.

- ✓ Change of geographical context and change of database:
  characteristics varying consequently, when the geographical scenario or the database features change in their turn, are too many compared to the available literature than we are not able to identify any linked recurrence.
- ✓ Change of the industrial sector:
  even in this case the rank of all the possible industrial sectors is too large and studies produced to date are not so numerous to allow for expressing any opinion. It would be interesting undertaking a wide systematic exam on such a problem.
- ✓ Change of the nature of agents:
  it seems that evolution of households/families' condition is not well represented by the Markov 1st-order stationarity, which seems in its turn to work a little better to describe firms' dynamics.
- ✓ Change of the variable:
  The most studies deny that quantities related to industrial production and incomes (both in Europe and in U.S.A.) evolve in compliance to the basic chain; debt and total assets (which have a financial nature), instead, appear to be satisfactorily approximated by the simplest Markov model. It should be investigated in more depth if there exist above all a correlation between the nature of the variable and the type of economic agent respect to the reliability of the chain.
- ✓ Dependence on the dynamical structure of the model:
  from the comparison of Adelman (1958) and Duncan-Lin (1972), on a side, with the remaining part of our bibliography treating Markov chains, on the other side, a scholar might deduce that credibility of the

basic chain respect to future economic distributional frequencies, at least in the short run, improves when individuals' entry-exit dynamics is somehow incorporated into the model, either by means of a simple, double-duty, input-output reservoir or by a stochastic model for accesses along with an inactive pool reaping all outgoing units.

- ✓ Dependence on the periodization in the statistical analysis:
  Shorrocks (1976) and Bickenbach-Bode (2003), not favourable to the time-independent 1st-order model, in the data processing make use of pluriannual intervals: it looks as if pluriannual analysis penalizes the model; the periodization in the data processing of both Hallberg (1969) and Duncan-Lin (1972) is made year by year: in the former the model is refused, in the latter the Markov first-order stationarity is confirmed.
- ✓ Dependence on numerical methods or approximations:
  it is plausible that, in the 1960s and early 1970s, some undisclosed numerical methods and approximating functions were expended for the elaboration of data, above all for resolution of chi-squared tests. We do not know the impact of using those methods and approximations on Markov applications.
- ✓ Errors in the data collection and elaboration:
  mistakes while data had been collected or during the import phase or in algorithms cannot be excluded; almost all publications about economic sciences never bear the computer codes whereby data were processed and tests were put in place, so such broad class of errors cannot be tracked down.

Chapter 3
"*MarkovInfer_MobInd.m*": a software for some classic mobility indices based on Markov chains

3.1 Introductive presentation of the device

As we have surveyed in the first introductive chapter, discrete-time stochastic processes defined over a discrete-state space, particularly discrete-time Markov chains, have been employed for decades for describing the dynamics of agents in a sample or a population respect to a relevant economic variable (very often it is income or wealth), founded on a range of k states labelled by the integers 1, 2, …, k, starting from an initial distribution at the initial time t_0. A consequent related issue is the measurement of the degree of mobility of the agents in the sample among the same states till a successive instant t_1.

As it will be seen in one of the following chapters, we have directed the attention to mobility as a scalar function $I(P)$ where $P = \{p_{ij}\ ;\ i,j = 1,2,..,\ k\ \}$ is the $k \times k$ transition matrix composed by the probabilities that an individual evolves, during the time interval $[t_0\ ,\ t_1]$, towards the final state j on condition that it has started from the state i. For details about the axiomatic definition and the theoretical framework of mobility indices established on transition matrices, see for example Shorrocks (1978).

As illustrated extensively in Fields (2000), Fields (2008), Fields and Ok (1996), Fields and Ok (1999), the function $I(\cdot)$ has to be chosen referring to the concept of mobility we are interested in (a more succinct review can be found in Ferretti (2012)). The pivotal idea of any mobility index, as a functional of the transition matrix, is that the distributional change, due to movements of some agents from a state i to another state j, is measured through the conditional probability $\{\ p_{ij}\ \}$ for any pair of states. Mobility indices are conceived as statistics summarizing the average mobility in the sample (population). Since any transition matrix is stochastic, that is every row must sum to 1, the index evaluation involves at most $k^2 - k$ independent terms (the

number of terms is quadratic respect to the number of states k). In addition, the calculus may get more complicated for the presence of arbitrary parameters and because the empirical transition probabilities $\{\hat{p}_{ij}\}$ are not immediately available and have to be estimated by means of statistical frequencies derived from an amount of data which must be as meaningful as possible, so data should be as large as possible, too. Therefore, the analysis of mobility is demanding under two aspects: the data archive and the automatic calculus. These are the reasons why the use of computer software is essential. Sometimes commercial programmes have been employed, surrounded by the company's secrecy; some other times software compiled just for the occasion are needed.

Incidentally, for a long time, data processing aimed at scientific publication, not only in statistical economics, has been committed to groups of collaborators of the main authors or to computational centres. Leaving mobility aside, often databases were expressly created during previous sophisticated research, involving many people for long. We have come across the old volume edited by Glass in 1954, referring to the interviews done for the 1949 British Social Survey and helpful for Prais; in the paper of Dunne, Roberts, Samuelson (1988) the characteristics of the firms' archive, devoted to this work, are exposed in depth by its authors; *ad hoc* data had been assembled for the study in Gort and Klepper (1982) and then the same data were used in Klepper and Graddy (1990).

Most of the empirical works concerning any mobility measures report in tables and graphs the computed values of the index under study, sometimes together with its statistical description, according to specific data collection, but almost none of them neither shows the software used for the analysis of data and the automatic calculation of that index, nor explain in detail how to compile it. Nowadays, the good computing power of personal devices allows a scientist, with sufficient programming skills, to analyse big databases on your own; international communities supporting the sharing of opensource, free software have been developing. However, there is still a high degree of reluctance to publish programmes used for research.

We propose a prototype of a free, open access, interactive software, called ***MarkovInfer_MobInd.m*** [1], which is a single piece m-file, so no additional files are required to work with it: its code is copied at the end of the present thesis, in Appendix A1.

It realizes inferential chi-squared tests on some features about Markov Chains, to be inferred from individual-by-individual data (equivalently, 'raw data' or 'granular data' or 'microdata'), and computes a set of some classic mobility indices, each of them is defined as a special functional of the transition matrix, both in the case of a first-order stationary Markov chain and for a not-stationary one. It operates as an automatic device, accessible to anyone without any informatic expertise, incorporating all the needed inputs, directly provided by the user, importing and processing by itself properly organised (as we are going to see in the continuation of the chapter) microdata, originally contained inside an xlsx or an xls-spreadsheet. Unbalanced datasheets filled by microdata, where thousands or tens of thousands of statistical units are archived, can be elaborated: each row must refer to a specific unit and contain its observed values for a scalar positive variable, relative to a sequence of observation dates. We shall also indicate them as 'microdata sheets' or 'granular data sheets' or 'longitudinal/granular/raw-panel sheets'. The number of observation dates registered in the file, that can be handled, is arbitrary if the spreadsheet is well-organized; the maximal set of datasheet's incumbents is soon extracted. By the phrases '*incumbent units*' or '*resident/permanent agents in the sheet*' or '*residing individuals*' (and other similar phrases) we shall denote, in the whole thesis, *all those statistical units inside the microdata sheet exhibiting a datum at every registered date for the current variable* (i.e. all individuals permanently appearing inside the sheet). A collection of quantitative outputs, structured as bi-dimensional or multidimensional arrays, is produced step by step, concerning the considered mobility measures over different time scales, and stored into the Matlab workspace. Some graphical outputs are created too, to weigh different indices up against one another for

[1] Copyright ©Danilo Aringhieri 2017. The software is released under the license Creative Commons Unported, Attribution - Not Commercial - Share Alike 4.0 (CC BY-NC-SA 4.0) whose readable summary is displayed at: http://creativecommons.org/licenses/by-nc-sa/4.0/ and whose legal text is available at: http://creativecommons.org/licenses/by-nc-sa/4.0/legalcode. At the beginning of the run, the same disclaimer will be displayed in the command windows and saved in the workspace with the name '*License*'.

the same underlying model and to study their dependence on time. They take a little more time to be produced than quantitative arrays.
It is a simple *command line interface* (CLI) for the MATLAB environment, exploiting basic commands and functions available before the release R2006a. The CLI has been chosen instead of the GUI (Grafical User Interface) because the former is lighter, though less fashionable too.
We outline that the tool is not designed to forecast, by a Markov approach, future distributions via past and current granular data for the variable in question, but it focuses on the statistical assessment of dependability for some Markov peculiarities and on the calculus of nine mobility indices based on the first-order transition matrices. Anyway, the fundamental parameters synthetizing single-agents movements within the longitudinal panel are reckoned (like the number $n_i(t)$ of agents jumping from any state i at time t; the count $n_{i\,j}\,(t)$ of one-step, or one-period, transitions from a state i, occupied at the date $t-1$, toward a state j occupied at the next date t; the number $n_{i\,j\,k}\,(t)$ of individuals jumping from a state i to a state j between $t-2$ and $t-1$, then jumping from the state j to a state k between $t-1$ and t). So, extensions are practicable to allow for predictions.
As introductory references about Matlab programming we suggest the various editions of publications by W. J. Palm III for engineers and scientists (edited by McGraw-Hill) along with the Mathworks online documentation at 'www.mathworks.com/support/'. Lots of other manuals are dedicated to Matlab applications for specific disciplines: about statistics, one of them is Martinez and Martinez (2007).
This thesis can also be considered as an interdisciplinary support, combined with the prototype *MarkovInfer_MobInd.m*, for researchers in economics and social sciences interested in inferential methods about Markov Chains and in mobility measures but without strong skills in software programming required to process big data.

The whole run of the software is divided in the following stages. When in doubt, its execution can be stopped at any time by pressing "Ctrl C" on the keyboard

1° At the beginning it provides and collects some preliminary information about the framework of the data by didactic paragraphs and "Yes or

Not" questions, then it requires the name (with the extension) of the xlsx or xls-type file, for initiating the automatic import and to extract the datasheet's incumbents.

2° After the automatic import, the data essential statistical structure is shown, i.e. the minimum value and the maximum one, the set of quantiles (the quantile order must be chosen by the user) as well as the vectors of the means, the medians and the modes for each observation period.

3° The tool requires the number of states, underlying the implemented Markov models, along with the points of the entire data-spectrum partition demarcating such classes. So, it is ready to count all the distribution frequencies of individuals among the states which are necessary for the *maximum likelihood estimations* (MLE) of the *first order Markov chain* conditional probabilities, both in the *stationary* case and in the *not-stationary* one, as well as the second order stationary probabilities.

4° Thereafter both the *chi-squared tes*t and the *maximum likelihood ratio criterion* are applied to two pairs of the type [null hypothesis, alternative]: the null of a stationary first-order Markov chain from the data against the time dependence of first-order Markov transition probabilities; the null that data are governed by a stationary first-order Markov chain against the fact that the stationary chain is second-order. They are two of the tests reported in Anderson and Goodman (1957), one of the fundamental works on statistical inference about just Markov chains. Moreover, the number of the system's degrees of freedom is reduced relative to its nominal value following the suggestion in Bickenbach and Bode (2003). At the end of this stage the average permanence times in every state are computed for a first-order chain, in the stationary case and in the not-stationary one, as well as through a third, purely mechanistic methodology. The results are displayed in the command window, while the programme is going on.

5° The tool asks for the weights $\{\omega_i\}_{i=1,\ldots,k}$ and the function of the intensity of jumps $v(|j-i|)$ the scholar prefers, characterizing particularly two indices, plus some other parameters contributing to the mathematical definitions of the mobility measures.

6° Then the list of well-known *mobility indices, which are functionals of the transition matrix*, are calculated *ceteris paribus* for the MLE first-order not-stationary Markov model, in the case of any one-period interval seen from the database, and for its MLE stationary analogue. Soon after, the *aggregated-in-time counterparts of these indices* are calculated, too, for each multi-period interval from the first observation time, inside the data, until everyone of the successive observations, for both the first-order Markov schemes. The mobility indices of our interest are: Trace and Determinant indices, the Second Greater Eigenvalue and the All Eigenvalues indices, the Index of Predictability, the Second Bartholomew Index, the Directional Index and the Mobility-as-Movement Family of measures, the Asymptotic Half-Life Index. All their values are arranged inside arrays, which will be stored into the Matlab workspace at the end of the run, together with the Markov transition matrices. These quantitative outputs are too many to display them all in the command window, but they can be unfolded after accessing the workspace.

7° At the end of the run some graphical outputs are also produced in separate windows. Some diagrams present all the one-period mobilities for the not-stationary first-order chain, in some other diagrams there are all the multi-period indices, both in the stationary first order Markov case and in the not-stationary one. These graphical outputs allow for comparisons, to study differences induced on the mobility measures by the two models and an insight into the different mobilities' trends over time.

Since *MarkovInfer_MobInd.m* adapts itself to any well-tailored xls/xlsx-type file, Markov models and mathematical expressions of the mobility measures, implemented in the tool, unavoidably depend on a large set of "free parameters", whose values and shapes have to be inserted by the user according to its preferences. Thus, we have attempted to make the software able to check the correctness of such entries respect to any related mathematical or logical constraint, and to a wide range of typing errors. If any input should break the constraints, the user will be advised and invited to correct his choice to go ahead with the run.

The decision to link our software to original data files formatted as '*.xls' or '*.xlsx' was taken after detecting some troubles regarding to the automatic data import from other file formats (like '*.txt') exported from the archive at our disposal. Besides, any modification to clean, simplify, correct the structure of the data (e.g. cancellation of columns or rows filled by strings, detection of units exhibiting incomplete time series) is easier in a spreadsheet than in a processor for 'txt' files.

Apropos the variable to work with, we have already mentioned the tool is planned to manage *quantitative scalar positive variable*, so the states of the transition matrix, constructed on the basis of the data, are numerical intervals; however, the definitions of mobility, we are interested in, are theoretically valid also for an *ordered qualitative variable* whose transition states are categories. This mismatch can be overcome provided that the user operates directly on the panel following this simple example. Considering a sample of firms, if they are classified as 'small', 'medium' or 'big', these three categories are replaceable (paying attention to their ordinal ranking) respectively with the discrete numbers '1', '2' and '3'; whereupon any real partition like { [1/2, 3/2); [3/2, 5/2); [5/2, 7/2) } or { [1, 2); [2, 3); [3, 4) } (where, for instance, [1/2, 3/2) includes '1', '2' is found in [3/2, 5/2) and '3' is in [5/2, 7/2)) may be entered during the run, to establish an ordered bijection between each original category and one of such numerical intervals.

3.2 Possible empirical investigations by means of "*MarkovInfer_MobInd.m*"

After the execution has been concluded, a great number of outcomes is saved in the Matlab workspace; most of them are bi-dimensional or multidimensional arrays, but there are also some graphics. The two models implemented in the script for almost all these outputs are the stationary and the not-stationary 1st order Markov chains. We have already recalled in the first chapter the early success the 1st-order stationarity met in economics before being carefully verified respect to more and more thorough data and tests, and then criticized in favour of either the 2nd-order stationarity or the 1st-order not-stationarity. The statistical inference on the former alternative is a little more complicated to be computed than for the latter; on the other side, previsions in the future, by

projecting a not-stationary 1st-order chain beyond the time horizon of the datasheet, demand in their turn a specific modelling. There are many options to realise not-stationary projections, among which we can instance the method proposed in Hallberg (1969), using multiple regression techniques and mentioned in the Chapter 2.

Albeit it has been long that the stationary 1st-order chain is often considered inadequate for distributional previsions beyond the short run in economics by a large community of scientists (*as proposed in Shorrocks* (1976) *and Schluter* (1997), *then probed in Bickenback and Bode* (2003) *for households' incomes, already summarised in the first chapter*), *we believe that one-period transition matrices coming in succession in a 1st-order, not-stationary chain are an efficacious tool to illustrate mobility in a sample or a population*.

So, the ideal employment of our software is, above all, executing comparative studies:

- ✓ between different databases for the same Markov model, the same variable, the same mobility index;
- ✓ between the different effects of the two Markov schemes on the measures of mobility and the related features, for the same datasheet and for a certain variable.

Such comparisons can be made for every one of a wide range of economic variables (in case of firms, from total assets to the number of employees, from total sales to the debt), to investigate how, and how much, the nature of the variable itself, not only the underlying theory, influences the measure of mobility.

Dealing with transition matrices, the dependence of the "clustering-on-the-main-diagonal" feature on the Markov approach and the variable could be inspected. Someone may wonder whether the 1st-order stationarity tends to underestimate/overestimate systematically, respect to the 1st-order not-stationarity, the mean permanence time in every class (and in some classes more than in the others), consequently overestimating/underestimating the mobility measures.

Another question is if the results of the two chi-squared hypothesis tests performed by *MarkovInfer_MobInd.m* do change when different variables for the same sample, or different industrial sectors for the same variable, are examined.

While observing graphs about one-period not-stationary indices over time, minima/maxima could be searched for the same dataset, asking for the following questions.

- Is a local-extreme behaviour exhibited only by few indices or by all indices and is it occurring over the same time interval?
- How, and how much, does the extreme behaviour of mobility change when the nature of the variable changes?
- Are all the considered indices effectively normalized into the real interval [0,1], or not?
- Does any index assume systematically too small values or unnecessarily large ones?

The software is also able to compute the aggregate-in-time, or multiperiod, versions of all the mobility measures of our interest, fixing the first observation date as the starting time, then considering each successive date as the ending time (the last graphical outcomes just refer to them).

- One of the questions is about what kind of trends are shown by the multiperiod mobility indices over time for the two Markov models, comprising underestimation (overestimation) phenomena.
- Does every economic variable exhibit its own peculiar mobility trend?
- Someone may examine whether the data confirms the conjecture, maintained by social scientists (e.g. Shorrocks (1978)) and already proved especially for earnings by some empirical studies (see Atkinson, Bourguignon and Morrisson (1992)), that any mobility measure increases when the timespan gets longer: in this regard, is there any dependence on the variable or any difference between the 1st-order stationary case and the not-stationary one?

3.3 Further developments of the software (and some digressions)

Our Matlab device might be developed in various directions.

Algorithms for detecting outliers among the data, to exclude them from computations, would refine the automatic import phase.

The issues of the subjacent theoretical model and the statistical estimation method, determining the transition matrix, could be faced with a more complex

approach. Beside a chain of the first order, stationary or not, it is possible to implement a second order Markov model, both in the classic form of a two-step composite transition among three successive simple states and in the alternative scheme as a first order chain between successive couples of composite states, every one of them made up of two simple classes. Chi-squared test and maximum-likelihood ratio criterion are also applicable to the null hypothesis of trustworthiness of a second order stationary chain against the alternative of a third order one and, more generally, to the null of a *u*th-order Markov process against the alternative it is of order *r*th but not *u*th. Next to simple Markov chains of a given order, some Markov mixtures could be placed, like the mover-stayer model, to investigate comparatively the impact of such models on each mobility index defined over the transition matrix. Conditional transition probabilities might be estimated by kernel methodologies instead of using the maximization of the likelihood function for inferential purposes.

Fixed the theoretical framework, with some minor modifications the software would be able to infer estimations of the statistical parameters only on the most part of the time-sequence inside the panel, to simulate a forecast in the short run about some subsequent frequency distributions, in accordance to the model. Then, simulations could be set against real frequencies in the final section of the same time-sequence by some goodness-of-fit tests, like the Pearson's chi-squared test.

Dealing with the refinement of the theory describing the sample dynamics, since the tool implicitly refers to a stationary sample where both the number and the identity of the statistical units are constant in time, it deserves a separate discussion the attempt to *incorporate* individuals' *entry-and-exit* phenomena respect to the initial sample into any Markov paradigm. It would make more exhaustive and realistic the distributional projections about the system, at least in the short run, and the quantitative analysis of mobility founded on transition matrices. In other words, after the starting observation period, some new agents may enter the system (e.g. an industry, when we turn to the subject of market structure) while some others may leave it, as time goes on: such kind of evolution entails an estimation of the transition matrix, as well as concepts and mathematical formulations of mobility, which are nontrivial. In this regard, not only a powerful theoretical model along with a serviceable software are necessary, but it is also essential a reliable database, where true

entrances, due to birth processes, as well as true departures, due to death processes, are scientifically registered.

Among the lessons taught by academics in the last decades, we specially reflected on two of them. The simplest solution was given for the first time by Adelman (1958) and consists in joining to the Markov system a *double duty repository* (one more state labelled as the *zero*th one) where potential entrants come from and where the outgoing individuals converge, indicated as input-output reservoir. Although such extension does not significantly complicate the frame of the model at all, it suffers the drawback that identifying a pool of potential entrants is technically helpful for the analysis of data, but it is also very difficult, if not often impossible, in practice. A quick remedy could be the assumption that the single repository initially holds N potential entrants (Adelman fixed N=100,000), whatever the destination after a one-period transition, but the choice of N has some effects, whose projections are examined by Stanton and Kettunen (1967). They showed that the proportion of entities falling in each of the active states does not depend on N but the actual number of entities in the active stages depends very critically on such a choice. This sensitivity to decreases as N increases. Moreover, they observed a decline of the number of active agents at equilibrium while increasing the number of potential entrants. This suggests a criticism: for a given observed number of entries into the system, a large N inferentially implies a small probability of entry so, for each departure from the system, the reservoir is very nearly absorbing. Conversely, a small N makes the reservoir reflecting. Nonetheless, in most practical situations, exiting units from the system do not have the same probabilities of acceding to active classes like the units never previously in the system.

The second solution was sketched in Duncan and Lin (1972) in a very concise manner (perhaps, too concise). A parallel passive state is placed next to each of the active Markov classes, where the statistical units are distributed, with the sole aim to absorb all those units going out. Simultaneously, the inconvenience of the potential entrants' pool is avoided by attributing a predetermined *stochastic nature to entrances phenomena*, more precisely by positing that new agents will access every active state according to a renowned (*multinomial*) probability law. On the other hand, in a similar scenario it becomes more laborious constructing and estimating with mathematical accuracy the

transition matrix in any Markov approach: it is feasible in case re-entries into the system does not occur and all passive states can be lumped into a *single absorbing repository for exit*.

Inter alia, it is conceptually debated in statistical economics whether, for understanding distributional changes and mobility, entrances and departures must be considered from a genuinely demographic viewpoint and, hence, whether they must be clearly distinguished from similar phenomena caused by mergers, break-ups and acquisitions (or re-denominations succeeding to false failures), frequently involving economic agents (firms, families), in case of additive and non-additive variables. Generally, the inclusion of entrants and exiting agents implies a revision of inferential procedures about any Markov paradigm and chi-squared-type tests, too. ***For all the above reasons, we have focused on the maximal subset of incumbents* (*we repeat they are all the permanently-present individuals inside the electronic sheet*)**.

In its entirety, the problem of *entry-exit dynamics* of statistical units, which is still an open challenge both in a theoretical perspective and in empirical research, is also intriguing relatively to some other topics in economic sciences, like the market structure in Industrial Organization (see, in this regard, the repeated citations in the review by Sutton (1997) and, as an example of dedicated empirical study, Dunne and Roberts and Samuelson (1988)).

While sophisticating the transitions' model or the estimation techniques in *MarkovInfer_MobInd.m*, an additional generalization is the extension of the list of the computed indices to all the mobility measures known in the scientific literature, that are at least about twenty, to give a complete description of their diverse behaviour respect to the same transition dynamics.

Chapter 4
Preliminaries on microdata primal structure, import and interactions

We did not want to compile a script involving exclusively some already fixed numerical quantities, to be directly modified wherever and whenever it had been necessary, what would have made mistakes more likely. Rather, we have attempted to realise a flexible device, capable of automatically adapting by itself both to the structural characteristics of the raw-data grid (namely, the array where microdata have been arranged or, equivalently, the configuration of the longitudinal panel where granular data are distributed) and to the preferences of the researcher.

The two Markov 1st-order models implemented in *MarkovInfer_MobInd.m*, as well as the mathematical definitions of some mobility indices computed by the software, are based on a heterogeneous assortment of parameters that can be distinguished in two categories:

- ✓ the ones detected by measuring the structure of current microdata by means of some Matlab function like `size(.)` (data-detected parameters);
- ✓ the free ones, which are adjusted by the command line interaction (user-adjusted free parameters).

All of them are defined in the script not as predetermined quantities but as variables, configured step by step at the right moment during the run, what allows for the adaptability we sought. In the case of user-adjusted parameters, some sections of the device have the task of verifying their accuracy.

In this chapter we are also going to illustrate the range of some difficulties and problematic situations, regarding primary archives, we encountered when compiling a sequence of instructions to import automatically and efficiently granular data into the Matlab workspace as a matrix able to be managed by subsequent automatic computations.

4.1 Free parameters and interactions

The "free parameters", to be inserted by the scholar at the waiting prompt in the command line, often are real scalars comprehending:

- the number of possible columns inside the spreadsheet filled with numbers which are not interesting data for the random variable the scholar wants to study;
- the degree of empirical quantiles (whether they are quartiles or quintiles or octiles or deciles, and so on) the user is interested in;
- the number of intervals splitting the spectrum of the random variable and constituting the Markov states;
- the points of the partition delimiting these intervals;
- the weights $\{\omega_i\}$ of the Markov starting states inside the two more sophisticated indices from the list of computed mobility measures (the Directional Index defined in Ferretti and Ganugi (2013), the Mobility-as-a-Movement Index proposed in Alcalde-Unzu, Ezcurra, Pascual (2006)), if the user does not accept the default weights;
- the value of two distinct parameters, one of them contributing to the definition of the Mobility-as-a-Movement Index, the other to the Exponentiated Determinant Index presented in Shorrocks (1978).

It is considered a free parameter also the functional shape $v(|\ i - j\ |)$ of the modulation for the jump $|\ i - j\ |$ from the i-th stratum to the j-th stratum which is shared (for comparison's purposes) by the Directional Index and the Mobility-as-a-Movement. A further input to be given is the complete name of the spreadsheet from which the data are going to be imported.

To set such free features, the command line interface communicates with the user in three different ways, all of them based on the function '`string_variable = input(`*`assertion at the prompt`*`,'s')`' or '`numeric_variable = input(`*`prompt assertion`*`)`':

a. the trivial "press Enter to continue",

b. second order branching interactions by polar questions like yes/no questions,

c. the request for an appropriate input at the waiting prompt: it may be the numerical value for a parameter, the analytical shape for a mathematical function (written respect to a subsidiary variable) or a string of characters.

We tried to make interactions of the software with the user as robust as possible respect to typing errors or the input of both something with a different nature from the expected one (e.g. a vector or a complex number instead of a real scalar) and a quantity not satisfying the mathematical constraints of the case (like ordering, the sign, the belonging to a specific range, the existence of finite and not-undetermined values of a function over a range etc.). Mistakes are recognised and the waiting prompt for a correct entry is displayed over and over in the command line until an acceptable input is typed on the keyboard. There is an example of rightness inspection of the chosen input in the last paragraph 4.4.
The interaction of the *a*-type is used several times in the script to pause the run of the software and is executed by means of:

```
response=input(' ','s');
while isempty(response)==0
disp(' Please, press the "Enter" button to START!')
response=input(' ','s');
end
```

exploiting that the command Enter, pressed on the keyboard, corresponds to an empty string of text characters for the variable `response` and it is the only for which it holds `isempty(response)==1`. So, the first line prepares the command window, by the `input` function, to receive a text string (as indicated by `'s'`) as entry, giving it the name `response`, and the while-loop will iterate the instructions preceding the `end` until `isempty(response)` is zero, that is until the input string is different from the empty one associated to the Enter command.
The interaction of the *b*-type, based on a yes-or-no question, is also repeated at every point where it is needed to decide between two distinct alternatives and is generally executed by means of:

```
response=input('   ','s');
while strcmp(response,{'Y','y','N','n'})==[0,0,0,0]
disp('   (You must answer exclusively "Y" or "N" !) ')
response=input(' ','s');
end
if sum(strcmp(response,{'N','n'}))==1
```

```
  statement1
elseif sum(strcmp(response,{'Y','y'}))==1
  statement2
end
```

In the initial while-loop the control statement '`strcmp(response,{'Y','y','N','n'})==[0,0,0,0]`' compares, by the function `strcmp(..,..)` and being the possible logical outcomes either 0 or 1, the input string attributed from the keyboard to the variable `response` to every one of the characters `'Y'`, `'y'`, `'N'` and `'n'`: until `response` does not coincide with any one of them, that is until the outcome for `strcmp(response,{'Y','y','N','n'})` is the vector `[0,0,0,0]`, the request '`response=input(' ','s');`' will be restated. Otherwise, an if-elseif command is executed: if `response` is worth `'N'` or `'n'`, the tool will implement *`statement1`*, but if `response` is worth `'Y'` or `'y'`, the alternative *`statement2`* will be implemented. *`statement1`* and *`statement2`* might also be very complex combinations of Matlab instructions.
Throughout the run of the tool, some instructional sentences or paragraph are displayed to give useful information and explanations.
Data have to be collected in a wide-format panel, not necessarily balanced, for a single scalar positive variable $X(t)$. Let observations on $X(t)$ be related to the set $D = \{t_1, t_2, \ldots, t_m\}$ of successive ordered dates, which correspond univocally, from some point onwards, to the columns in the spreadsheet. During the run these dates will be identified by positive integers $\{1, 2, \ldots, m\}$ representing their position in the chronological ordering.

4.2 The automatic data import and the "soft cleaning"

A brief summary appears just after the launch, containing some clarifications about the original file's needed structure that permits the correct automatic import. *It has to be a xls-type or xlsx-type single spreadsheet*, exported by most data archives and very easy to manipulate: it will be enough inserting the complete file name, including its extension, when requested at the prompt, to import the data by the simple combination of instructions '`filename=input(.,'s');`' and '`[AuxData1,Titles]=xlsread(filename)`' in the script. By the former instruction, the name of the spreadsheet is placed as

first argument inside the '`input`' function, acquiring it as a string variable (according to the second argument `'s'`) named '`filename`'; by the function '`xlsread`' in the latter phrase, spreadsheet's contents are imported, subdivided in two arrays, and saved in the workspace. These two arrays are `AuxData1`, that shows all the numerical values from the sheet, filling each string cell's position with the '`Not-a-Number (NaN)`' acronym; and `Titles` that consists of all the string-formatted xlsx-cells, filling each numerical cell's position with a pair of quotation marks. In both arrays the mutual positions between cells are preserved. If the xlsx (or xls) file is composed by two or more spreadsheet containing columns filled with data, only data inside the first spreadsheet will be automatically imported. If such first sheet is not organised by columns and rows filled by data in numerical format, the automatic import fails. Inside the sheet each cell has to contain only one scalar value in numerical format to make the import in Matlab of this cell feasible. If Microsoft Excel is not installed on the computer but there is any spreadsheet able to read and save *.xls or *.xlsx files (CALC from the LibreOffice Suite, for instance), data are imported by means of the Matlab basic mode.

To compile a script which automatically imports data, it is better to refer to a specific database characterized by some basic features that may be different from the traits of other similar databases. For the present device we have referred to the archive AIDA, producing a great variety of datasheet structures (for a general presentation, see the webpage https://www.bvdinfo.com/en-gb/our-products/data/national/aida#secondaryMenuAnchor1). Studying spreadsheets exported from AIDA and the results of their import into the Matlab workspace, we found three kinds of sequence of cells that could get some problems (List [A]):

1° columns formatted cell-by-cell to contain only strings, among other columns wholly formatted to contain only numbers;

2° columns formatted cell-by-cell to contain only numbers, but which are not values of the variable $X(t)$ the user is interested in;

3° string cells filled with the acronym 'n.a.' or with any other string meaning 'not available (value)', surrounded by other numerical cells filled with interesting data.

Thus, we wrote the programme thinking about an ideal framework for the raw data, optimizing the import, represented in the following Table 4.1, where the

variable $X(t)$ is the number of employees annually recorded. The archive AIDA makes available, for all the accounted variables, quantitative information about the last ten years before the present one. Thanks to the ease with which electronic spreadsheets can be handled, the first two kind of problematic columns could be preliminarily sought and cancelled before using Matlab.

Name	Employees, Year 2010	Employees, Year 2011	Employees, Year 2012	Employees, Year 2013
Name1	194	119	131	150
Name2	4	n.a.	5	9
Name3	210	199	0	0
Name4	n.a.	10	n.a.	13
Name5	91	95	110	114

Table 4.1: a brief example of the simplest data structure automatically and directly importable in Matlab by our device

However, the tool tolerates some deviations, represented in Table 4.2 (whose content is totally invented) from the standard above, because it is capable to clean those flaws from the data matrix, before saving it into the workspace, to lead the data frame to the standard one.

Name	Ranking (numerical format)	SIC 2007 (text format)	Address	Zip code (numerical format)	Employees, Year 2010	Employees, Year 2011	Employees, Year 2012
Name1	1	360000	Sunset Boulevard, 14	55416	124	67	178
Name2	2	350000	Dawn Street, 9	55417	n.a.	119	n.a.
Name3	3	370000	Golden Avenue, 57	55416	210	199	0
Name4	4	370000	Upper Street, 1	55416	n.a.	4	6
Name5	5	360000	Champions Road, 45	55413	3	0	0
Name6	6	370000	Carlson Avenue, 82	55419	237	53	186
Name7	7	360000	Arlington Road, 7	55416	0	0	n.a.

Table 4.2: an example of a more complicated data structure, which our device is able to import automatically In Matlab after the 'soft cleaning'

We have called such removal of certain flaws, together with the extraction of the datasheet incumbents (see the final part of paragraph 4.2), as '*soft cleaning*'. More precisely, it may happen that electronic spreadsheets, exported from data archives about companies, often present some opening columns (see the first one, labelled as 'Name', in Table 4.2) or end columns, as well as some

rows on the top (the first in Table 4.2) or on the bottom, filled only with string characters representing the names of all firms, their addresses or different identification codes registered as strings (postal or fiscal codes, VAT numbers and similar), dates or variables' names and so on. Matlab simply ignores these surrounding string columns and string rows during the import, so they are not annoying at all.

Nevertheless, sometime columns made of strings might be intermixed with the columns filled with true data (e.g. the third, labelled as 'SIC 2007' and formatted as a text, and the fourth containing addresses, in Table 4.2): it is the 1st complication indicated in the List [A] of this chapter. They must be treated with some attention since they are interpreted as columns, consisting of '`NaN`' outcomes, between other columns entirely composed of true numbers (see the second and the fifth, as well as the last three ones about the number of employees, in Table 4.2). *MarkovInfer_MobInd.m* locates inside `AuxData1` such interior string columns and register their position into the vector `nan_col` by the following for-loop:

```
for k=1:dim_AuxData1(2)
nan_col(k)=(sum(isnan(AuxData1(:,k))))./dim_AuxData1(1);
end
```

`nan_col` is a row vector whose *k*-th component is worth 1 if and only if the *k*-th column in `AuxData1` is a string column, is worth 0 or is positive but less than 1 otherwise (i.e. if the *k*-th column contains at least one numerical cell).

Then the instruction '`not_nan_col=find(nan_col-1);`' selects all the columns inside `AuxData1` containing at least one numerical cell and by '`AuxData2=AuxData1(:,[not_nan_col]);`' all string columns are left out from the new auxiliary array `AuxData2`, forerunning the clean data matrix (named `MatData` in the script).

A more troublesome kind of columns in datasheets are the ones, cited at the 2nd point in the List [A], full of numbers which are not values of the process $X(t)$ at all. Normally they are identification codes (e.g. the zip codes formatted as numbers in the fifth column in Table 4.2) or a column of ranking positive integers which correspond one by one to every individual according to some order criterion (the second column, labelled as 'Ranking' and in numerical

format, too). Sometimes it happens to find in datasheets like the AIDA's ones, alongside the values of the interesting variable, some columns filled with true values of another variable that we do not care of and we would exclude, at least for the moment. Such non-data columns in numerical format, if not deleted or not converted in a text format in advance, will inevitably falsify any automatic inference from microdata about the transition probabilities and, consequently, the computation of any index based on transition matrices.

By the phrase '`nondata_col=input(' ');`', the *program demands*, and saves into the scalar variable `nondata_col`, the *number of these numerical non-data columns*. The instruction '`All_Data = AuxData2(:,nondata_col+1:dim_notnan_col(2));`' is able to *omit* them *from the true data matrix* `All_Data` *solely when they are all on the left of the values of the process X*(*t*), *as it happens for the second and the fifth columns in the Table 4.2*, *just on the left of the annually-recorded number of employees*. Unfortunately, the tool cannot keep out non-data numbers in any case, wherever they are in the primitive file: therefore, it would be better preliminarily to erase them, or to convert their format in a text-type, before processing the granular data file.

Anyhow the resulting `All_Data` in the workspace is a matrix where each column corresponds to one and only one of the periods registered in the datasheet, each row matches one and only one of individuals with a reported value at least for one period. The function `size(All_Data)` generates a bi-dimensional row vector `dim_All_Data` whose components are respectively the number of all registered individuals and the number of all the observed periods (for AIDA, such periods are the last ten years preceding the current one).

Finally, let us move on to consider the 3rd kind of problematic cells in the List [A]: *often* it happens that *longitudinal panels* concerning real samples *are unbalanced*, namely some statistical units may lack of a datum relative to a certain variable for some dates among all the registered ones. In this situation the corresponding cell in the sheet is filled with some strings (for instance, the acronym 'n.a.', meaning 'not available', or a simple dash) denoting such unavailability; the elements inside the imported matrix stored in the workspace, corresponding to every one of the cells lacking the datum, appear as a single

`NaN` outcome. Dealing with firms, this unbalancing problem of panel data, as the time passes, may occur for some reasons to be distinguished from each other:

- birth-death dynamics inside the sample;
- communication faults to the archive manager (regarding the Italian archive AIDA, some firms could decide not to convey information to the consortium InfoCamere relatively to some balance-sheet items for some periods);
- merger/acquisition/breakup phenomena involving sample's existing firms;
- false failures followed by changes of name.

Birth-death dynamics consists of events like either the definitive disappearance of an agent at a specific date (e.g. a failed firm which has left the market at a precise year), remaining without any datum from such date onwards; or the appearance of a new agent which has not been previously existing (e.g. the creation of a new business at a specific year), exhibiting a datum from such date onwards. The choice of taking into consideration birth and death events obviously makes not stationary the study sample and this entails a greater sophistication both in the definition of the system's set of states and of the estimation of the transition matrix, whatever the scheme adopted, than when the sample is stationary (incidentally, it might be necessary also to modify the concept of mobility the specific study is referring to). In addition to failures of existing businesses, which exit from the market, and entrances in an industry of newly established companies, the puzzle of integrating firms' demographic dynamics into any Markov paradigm or any quantitative description of mobility is made more difficult, than the description of the demographic changes of a human population, by the facts that two or more firms can merge into a single one or a more successful company can acquire another weaker one. There is no clear consensus on how to interpret a merger: whether it is the disappearance of old merging subjects and the simultaneous onset of a new one, as they were independent from each other, or not. Neither it is clear the interpretation of an acquisition: whether it is the death of the weaker company, independently from the contemporaneous growth of the stronger purchaser, or not. There again, accurately identifying all the previous demographic variations inside a non-stationary group may require an appropriate preliminary matching of multiple features to build an adequate spreadsheet, ever since the initial use

of the database. Moreover, it is now evident that demographic variations are not to be confused with the administrative issue of missing information caused by nonfeasance to the archive.

For the reason above, to avoid setbacks, *we decided to begin extracting the subsample of the spreadsheet incumbents*, defined as the ensemble of agents exhibiting a value for $X(t)$, at every period accounted for inside the sheet (the same thing has been done in Bertoni, Aletti et Alii (2018)). These values are saved into the Matlab workspace through the matrix `incumbData`. Such incumbents will be also denoted as 'the core of all incumbents', 'all incumbents' generation' or 'dataset residents', 'dataset permanent/residing/residential units', and other similar phrases.

Besides, from AIDA (the electronic archive of all Italian firms we were availed of) it is feasible to export datasheets for the almost-complete sub-population of businesses matching any fixed set of selection criteria. Separating the residential agents from the original datasheet gives the practical advantage of focusing on a more specified sub-population, whose size goes from 2000 to 6000 residents in the case of the five traditional Italian industries at a national scale (see Chapter 6). A size scale of few thousands is a compromise aimed to prevent two kinds of inconveniences without giving up a profound inferential estimation. The two inconveniences affecting significance tests, like the χ^2-type ones executed in Chapter 7 about some Markov features, are: the potential sensitivity to little deviations inside very large panels (e.g., a size scale of some hundreds of thousands; see Bergh (2015) and Bertoni, Aletti et Alii (2018)); the alleged tendency to refuse any null hypothesis (dreaded in the statisticians' community) in the presence of big amounts of data. Parenthetically, the longitudinal core of all persisting units can be viewed as a cross-sectional panel over the registered decade, where all singles are bringing back up at every date (see paragraph 1.1 in Chapter 1), hence it is clearly a degenerate cross-sectional ensemble on an extended epoch. It is advisable and more correct to apply Markov inferential techniques thought up in Anderson and Goodman (1957) for a fixed population evolving with the time on well-known paths, instead of any cross-sectional approach (not skilful in capturing the progress of a stochastic chain), akin to the ones in Bergh (2015) itself.

The extraction is made from the wider original sample of all individuals expressing at least one value for $X(t)$ during the historical sequence, which is in its turn imported into the workspace in the form of the matrix `All_Data`. Operations are carried out by the lines explained below.

```
dim_All_Data=size(All_Data);
for k=1:dim_All_Data(1)
incumbent_row(k)=prod(isfinite(All_Data(k,:)));
end
incumbent_posit=find(incumbent_row);
incumbData=All_Data([incumbent_posit],:);
dim_incumb_pos=size(incumbent_posit);
num_of_incumb=dim_incumb_pos(2);
dim_incumb=size(incumbData);
```

The for-loop determines the row vector `incumbent_row` whose components are as many as the number of all agents registered and are worth 0 if, and only if, the agent corresponding to the vector component lacks the datum for at least one observed period, or 1 if, and only if, the agent is a datasheet resident. Indeed, `All_Data(k,:)` fixes the entire k-th row in the matrix of all individuals; while the function `isfinite(All_Data(k,:))` assigns for every column (i.e. for every observed period) the value 0 exclusively to a `NaN` element in the k-th row, the 1 to a row position occupied by a true finite number; then the phrase '`incumbent_row(k)=prod(isfinite(All_Data(k,:)));`' reckons the product of all the 1 entries together with eventual 0 entries in the fixed k-th row: evidently the product `incumbent_row(k)` is 1 if, and only if, the k-th row is appointed to a resident (lacking a `NaN` entry for every period, by definition). Finally, `find(incumbent_row)` search for the linear positions of nonzero elements of `incumbent_row`, that are equivalent to the positive integers identifying residing singles in the primitive raw-data grid, and collects them in progressive order into a new row vector `incumbent_posit`. The number of its component '`num_of_incumb=dim_incumb_pos(2);`' is exactly the number of incumbents. The matrix of the permanent data for $X(t)$, during the overall recorded age, is extracted from `All_Data` and stored into the workspace by '`incumbData=All_Data([incumbent_posit],:);`'.

4.3 Preparing the estimation of the transition probabilities

After the automatic import, both the clean matrices `All_Data` and `incumbData`, containing only the observations about the process $X(t)$, are saved into the workspace. For instance, if the primitive granular-data file included the extract written in Table 4.1, the related extracts in `All_Data` and `incumbData` would be as represented respectively in Table 4.3 and Table 4.4

194	119	131	150
4	n.a.	5	9
210	199	0	0
n.a.	10	n.a.	13
91	95	110	114

Table 4.3: the result of the automatic import, by MarkovInfer_MobInd.m, into the Matlab workspace, of the little datasheet in Table 4.1

194	119	131	150
210	199	0	0
91	95	110	114

Table 4.4: the result of the residents' pulling-out from the imported matrix of all individuals in Table 4.3

where each column corresponds chronologically to one of the years from 2010 to 2013; inside `All_Data` each row corresponds to one and only one of all the statistical units, but in `incumbData` each row corresponds to one and only one of those individuals expressing a numerical datum for every observed year. Note, in the sole `All_Data`, the `NaN` values in the positions of all the original cells filled with a 'not-available-datum' string. Whatever the original file, the instruction `size(..)` measures dimensions of `All_Data` and `incumbData` to build two bi-dimensional vectors `dim_All_Data` and `dim_incumb` in the script: by construction, the row-countings `dim_All_Data(1)` and `dim_incumb(1)= num_of_incumb` are respectively the number of all the recorded agents and the number N of all recorded residing singles respect to $X(t)$; the number of columns `dim_All_Data(2)= dim_incumb(2)` is the number m of dates when $X(t)$ is measured in the sample (for AIDA it is constantly m=10, and the consecutive ten years shift forward once a year). The

components of `dim_incumb` are fundamental parameters, repeatedly used in computations.

Then the device detects (by some operators from the Matlab Statistical Toolbox) and displays (by the basic functions `disp(output)` or `fprintf(..)`) the statistical structure of the current datasheet, to allow the scholar for comprehending how to properly split the whole spectrum of the random variable $X(t)$ into the subintervals intended to become the Markov states among which the panel's units move. The following related outputs will not canceled from the Matlab workspace, to remain available to the researcher's considerations. By means of '`MaxValues = max(incumbData)`', '`MinValues = min(incumbData)`', '`Means = mean(incumbData)`', '`Medians=median(incumbData)`' and '`Modes=mode(incumbData)`' five, *m*-component, row vectors are built up, containing, for each of the *m* observation dates, respectively the maximal and the minimal value of the variable, the sample's mean, median and mode; '`lowest_datum=min(MinValues')`' and '`highest_datum=max(MaxValues')`' find the smallest value of $X(t)$ and the greatest one inside the datasheet, whatever the date. Moreover, the user is asked (*b*-type interaction) whether he is interested in knowing, for every period, any particular set of empirical quantiles relative to the residents' group for the variable: if so, the degree of interest (`qntl_degree`, i.e. 4 for quartiles, 5 for quintiles, 6 for sextiles, 10 for deciles and so on), is requested (*c*-type interaction), the inserted scalar is checked and the array of such quantiles (`Quantiles`) is computed by the code:

```
qntl_degree=input(' ');
q=zeros(1,qntl_degree);
for k=1:qntl_degree
q(k)=k/qntl_degree;
end
Quantiles=(quantile(incumbData,q,1))';
```

The component '`q(k)=k/qntl_degree`' in the for-loop defines, being *k* an integer from 1 to `qntl_degree`, the cumulative distributional frequency, associated to the *k*-th position of every quantile inside its ordered set, as a multiple of the fraction of the normalized total probability. The instruction

`quantile(incumbData,q,1)` returns, for each column of the matrix `incumbData` (the third argument '1' indicates that, inside the scanned matrix, the column is fixed while the row is varying), a row vector of the same length of `q`. So, due to the transposition by the apex operator ' ' ', dimensions of the array `Quantiles` are *m*×`qntl_degree`, all the rows corresponds in a bijective way to the registered observation periods, its elements may be used as input for the successive interactions. Specifically, the set {`Quantiles(t,1:qntl_degree)`} at the date *t* could be adopted as the partition subdividing the spectrum of *X*(*t*) into the Markov classes (see the next chapter): the *k*-th class is delimited by the (*k*−1)-th and the *k*-th quantile. This choice would avoid, for the definition of quantile itself, the situation where at least one class is empty at the beginning of the transition from time *t* to time *t* +1, generating at least one column filled with zeros in the 1st order transition matrix, what may give problems for the estimation of some mobility indices derived from the spectral analysis of that matrix. Besides, the comparison between two Markov transition matrices, describing two completely different evolution processes occurring in a certain population, based on the same quantile set, instead of absolute intervals, is useful for the analysis of mobility concerned with the re-rankings of individuals through time: it avoids the apparent paradox underlined in Fields and Ok (1999), at section 2.5 (pages 566-567, transformations X-XI), arising when the number of agents in the respective classes of the two mobility processes is radically diverse.

4.4 Checking the correctness of a positive scalar input: the example of 'qntl_degree'

In this paragraph we present an example of Matlab audit code on the correctness of an integer, positive scalar input, like `qntl_degree`, provided by the user into the command line, at a *c*-type interaction. The simple mathematical traits of the parameter delineate a brief case list on which the checking is based because it is negated by any mistaken input. The extract from *MarkovInfer_MobInd.m* is constituted by two if-elseif operators nested with a while-loop (the numbers here identifying the lines are for convenience sake and they are not the same in the final version of the whole software):

```
1  qntl_degree=input('    ');
```

```
 2  if(isscalar(qntl_degree)==0 || isreal(qntl_degree)==0 ||..
    ...isnan(qntl_degree)==1 || isinf(qntl_degree)==1 ||..
    ...qntl_degree<=0 || floor(qntl_degree)~=qntl_degree)
 3  while(isscalar(qntl_degree)==0||isreal(qntl_degree)==0||..
    ...isnan(qntl_degree)==1 || isinf(qntl_degree)==1 ||..
    ...qntl_degree<=0 || floor(qntl_degree)~=qntl_degree)
 4  if isscalar(qntl_degree)==0
 5  disp(' ( You pressed the [ENTER] button or inserted a..
    ...vector or an array ! ) ')
 6  disp('   BUT  you must insert exclusively a positive..
    ...integer ! ) ')
 7  qntl_degree=input('   ');
 8  elseif(isscalar(qntl_degree)==1 &&(isnan(qntl_degree)==1..
    ...||isinf(qntl_degree)==1 || isreal(qntl_degree)==0))
 9  disp(' Your input is a NaN value or infinite or..
    ...a complex number: WHY? ')
10  qntl_degree=input('   ');
11  elseif qntl_degree<=0
12  disp('( You must insert exclusively a positive number !)')
13  qntl_degree=input('   ');
14  elseif (qntl_degree>0 && floor(qntl_degree)~=qntl_degree)
15  disp('(You must insert exclusively a positive integer !)')
16  qntl_degree=input('   ');
17  end
18  end
19  elseif(isscalar(qntl_degree)==1&& isreal(qntl_degree)==1..
    ...&& isnan(qntl_degree)==0 && isinf(qntl_degree)==0 &&..
    ...qntl_degree>0 && floor(qntl_degree)==qntl_degree)
20  end
```

Its chassis is the compound conditional operator if-elseif-end at the lines n° 2-19-20. The sentence at the line n° 2 ‘ `isscalar(qntl_degree)==0||isreal(qntl_degree)==0||isnan(qntl_degree)==1||isinf(qntl_degree)==1 || qntl_degree<=0 || floor(qntl_degree)~=qntl_degree` ’, after the if-conjunction, itemises all the error cases about the parameter, related to one another by the logical exclusive disjunction ‘*or*’ always represented as a double bar ‘ `||` ’. Its logical complement is the phrase occupying the line n° 19 after the elseif-conjunction, where are cited the fundamental features of the parameter, connected to one another by the logical conjunction ‘*and*’ always represented as ‘`&&`’. The fact that `qntl_degree` must be a real scalar is indicated respectively by `isreal(qntl_degree)==1` and `isscalar(qntl_degree)==1`, which also rejects a string, as well as any at least bi-dimensional vector or array, and any complex number, inserted as a value of the quantile degree. `isnan(qntl_degree)==0 && isinf(qntl_degree)==0` respectively mean that is not acceptable, as a

quantile degree, neither a `Not-a-Number` entry, like the undetermined ratio 0/0, nor any infinite quantity, i.e. the `Inf` input itself or any ratio where 0 divides a not-null number. Note the application of Boolean functions of the form `isproperty(Y)`: if the object `Y` satisfies the property mentioned inside the function name, the output is '`==1`', but, when the property is not verified, the result is '`==0`'. The request `qntl_degree>0` is obvious and the final `floor(qntl_degree)==qntl_degree` demands the parameter is an integer, having to coincide with its floor function. The error cases making up the sentence at line n° 2 (reported at line n° 3 after the conjunction 'while', too) clearly are, one by one, the negations of the previous conditions and the halfway while-loop at line n° 3-18 repeatedly both enables the control on the input respect to the error cases' sentence and proposes the specific warnings for every one of them (lines n° 5,6,9,12,15).

Mutatis mutandis, owing to the mathematical bounds of each input, the audit code for `qntl_degree` is at least very similar to codes for checking scalar inputs, like `nondata_col` at paragraph 4.2 or some other parameters we will mentioned in the following chapters (the `classes_num` needed to construct the Markov strata, `alfa` and `bet` inside two of the computed mobility indices).

Chapter 5
Construction and checking of the Markov space from granular data

5.1 Construction of the state space

During the historical period [0,T] of our interest, let us consider a set $D \equiv \{t_1, t_2, \ldots, t_r\}$ of successive ordered dates $0 \le t_1 < t_2 < \ldots < t_r \le T$, which corresponds one by one to the set of positive integers {1, 2, …, r} identifying their chronological position. In many practical situations, these dates are often equidistant, too. At every date, for a fixed not-negative real variable $X(t)$, the value expressed by each individual belonging to a stationary sample of N statistical units has been observed: sample stationarity means both N and individual identities are remaining unvaried in [0,T]. For example, the group of all spreadsheet incumbents, extracted by *MarkovInfer_MobInd.m* and mentioned in the previous chapter, is stationary. *A priori*, $X(t)$ may be a proper functional of any firm balance sheet variable (total assets, total sales, debts) or any firm size variable (outputs, number of dependent), as well as to households incomes, in a certain country or region. Anyhow, its value for the l-th sample unit, at time τ will be indicated as $\{X_l(\tau)\ ;\ l = 1, 2,.., N \text{ and } D \ni \tau\}$: all these values belong to the empirical spectrum $[\min_{\forall \tau:\, D \ni \tau,\ \forall l=1,\, 2,..,\, N} X_l(t),\ \max_{\forall \tau:\, D \ni \tau,\ \forall l=1,\, 2,..,\, N} X_l(t)]$, derived from the current spreadsheet, that can be virtually extended to the real halfline [0,+∞). Remember, from Chapter 4, that $\min_{\forall t,\forall l} X_l(t)$ = `lowest_datum = min((min(incumbData))')` and $\max_{\forall t,\forall l} X_l(t)$ = `highest_datum = max((max(incumbData))')` have already been computed by the tool at the stage we are considering. Let us consider also a set Q of possible values of $X(t)$, $Q \equiv \{a_1, a_2, \ldots, a_{m-1}\}$, ranked according to $0 \equiv a_0 \le \min_{\forall t,\forall l} X_l(t) < a_1 < a_2 < \ldots < a_{m-1} < \max_{\forall t,\forall l} X_l(t) < a_m \equiv +\infty$ in such a way that the rightside real halfline can be splitted as the union of intervals of the partition $S_1 \equiv \{\ [a_j, a_{j+1}),\ j = 0, 1, 2, ..., m-1\}$. This partition remains fixed in

time, like the sample. In order to implement any Markov chain model to data analysis, each interval of the partition is to be identified with every one of the m states on which the chain is based, either it is not-stationary or stationary. So, after the configuration and the execution of the automatic, raw-data import, the user must choose how to shape the structure of the state space, that is he must provide into the command line, by a *c*-type interaction (see Chapter 4), the number of intervals m and the points of Q delimiting such classes inside the spectrum of $X(t)$, to go ahead in the run. The positive integer m is named `classes_num` in the software's text and the correctness of its inserted value is verified by the same code, explained in the last paragraph of the previous chapter, used for the input `qntl_degree`, that is neither not-scalar quantities nor not-positive ones, neither infinite-undetermined entries nor string formatted ones are accepted. The partitioning points of Q in the script are the positive scalars `sup(1)`$\equiv a_1$, `sup(2)`$\equiv a_2$, …, `sup(classes_num-1)`$\equiv a_{m-1}$, which are controlled also respect to their aforementioned ranking inside the data spectrum. Then they are collected in the row vector `sup`, used to create the (m+1)-component vector '`partition=[0,sup,Inf]`', the base for implementing the statistical inference on Markov classes from the current spreadsheet.

5.2 Checking the states' bounds

The Matlab audit code on the correctness of inputs for the component of `sup`, the partitioning points' set, is a combination of two variants of the analogue code for an integer, positive scalar input. It is an application of the case list just below, consisting of the simple mathematical properties of a good partition: any mistake on a_j is the logical negation of at least one among these properties.

i. $a_j > \min_{\forall t,\forall l} X_l(t) \quad \forall j = 1, 2, \ldots, m-1$

ii. $a_j < \max_{\forall t,\forall l} X_l(t) \quad \forall j = 1, 2, \ldots, m-1$

iii. $a_{j-1} < a_j \quad \forall j = 2, 3, \ldots, m-1$

iv. $a_j \neq$ (∞ and NaN) and is not a complex number $\forall j = 1, 2, \ldots, m-1$

v. a_j is neither a vector nor an array $\forall j = 1, 2, \ldots, m-1$

During the sequence of *c*-type interactions, the *iii* property is the only one used to test any current input respect to the previous ones, thus it is not applicable to

the first point a_1 along with the other four properties, and at the checking it is necessary to separate a_1 from the successive points. The concept of the code is:

an external for-loop restates the request for right entries about the partition;
a) if the for-loop is executing the 1st step:
1) if the compound condition (*i* and *ii* and *iv* and *v*) is not satisfied
(the simple condition *iii* must be omitted):
the request for a correct input is iterated, until the input is wrong,
together with a message concerning the possible mistake,
2) otherwise the input is accepted and saved into the workspace;
b) otherwise, if the for-loop is executing any one of the steps after the 1st:
1) if the compound condition (*i* and *ii* and *iii* and *iv* and *v*) is not satisfied
(the simple condition *iii* must be included):
the request for a correct input is iterated, until the input is wrong,
together with a message concerning the possible mistake,
2) otherwise the input is accepted and saved into the workspace;
the for-loop is ended when the last input for the partition is also right.

The text of the audit code is (lines' numeration may not be the same in the definitive script):

```
315 for k=1:classes_num-1
316 fprintf('Insert the top (expressed as number of ''%s'')...
    ...of the %-1.0d° class:\n',units,k)
317 sup_k=input(' ');
318 if k==1
319   if not(isscalar(sup_k)==1 && isreal(sup_k)==1 &&...
      ...isnan(sup_k)==0 && isinf(sup_k)==0 && sup_k>0...
      ...&& sup_k>lowest_datum && sup_k<highest_datum)
320   while not(isscalar(sup_k)==1 && isreal(sup_k)==1 &&...
      ...isnan(sup_k)==0 && isinf(sup_k)==0 && sup_k>0...
      ...&& sup_k>lowest_datum && sup_k<highest_datum)
321     if ( isscalar(sup_k)==0)
322   disp(' ( You pressed the [ENTER] button or inserted ...
      ...a vector or an array ! ) ')
323   disp(' BUT you must insert exclusively a positive ...
      ...integer! ) ')
```

```
324   sup_k=input(' ');
325     elseif (isscalar(sup_k)==1 && (isnan(sup_k)==1 ||...
      ...isinf(sup_k)==1 || isreal(sup_k)==0))
326   disp(' Your input is a NaN value or infinite ...
      ...or a complex number: WHY? ')
327   sup_k=input(' ');
328     elseif sup_k<=0
329   disp(' ( You cannot insert a positive number or zero! )')
330   sup_k=input(' ');
331     elseif sup_k<=lowest_datum
332   disp(char({' Insert a value bigger than Minimal datum ...
      ...(but smaller than Maximal datum): '}))
333   sup_k=input(' ');
334     elseif sup_k>=highest_datum
335   disp(char({' Insert a value smaller than Maximal ...
      ...datum (and bigger than Minimal datum): '}))
336   sup_k=input(' ');
337     end
338   end
339   else
340   end
341   sup(1)=sup_k;
342 elseif k>1
343   if not(isscalar(sup_k)==1 && isreal(sup_k)==1 && ...
      ...isnan(sup_k)==0 && isinf(sup_k)==0 && sup_k>0 ...
      ...&& sup_k>lowest_datum && sup_k<highest_datum ...
      ...&& sup(k-1)<sup_k)
344   while not(isscalar(sup_k)==1 && isreal(sup_k)==1 && ...
      ...isnan(sup_k)==0 && isinf(sup_k)==0 && sup_k>0 ...
      ...&& sup_k>lowest_datum && sup_k<highest_datum &&...
      ...sup(k-1)<sup_k)
345     if (isscalar(sup_k)==0)
346   disp(' ( You pressed the [ENTER] button or inserted ...
      ...a vector or an array ! ) ')
347   disp(' BUT you must insert exclusively a positive ...
      ...integer ! ) ')
348   sup_k=input(' ');
349     elseif (isscalar(sup_k)==1 && (isnan(sup_k)==1 || ...
      ...isinf(sup_k)==1 || isreal(sup_k)==0))
350   disp(' Your input is a NaN value or infinite or ...
      ...a complex number: WHY? ')
351   sup_k=input(' ');
```

```
352     elseif sup_k<=0
353   disp(' ( You cannot insert a positive number or zero! )')
354   sup_k=input(' ');
355     elseif sup_k<=lowest_datum
356   disp(char({' Insert a value bigger than Minimal datum...
      ...(but smaller than Maximal datum): '}))
357   sup_k=input(' ');
358     elseif sup_k>=highest_datum
359   disp(char({' Insert a value smaller than Maximal datum...
      ...(and bigger than Minimal datum): '}))
360   sup_k=input(' ');
361     elseif sup_k<=sup(k-1)
362   disp(char({' Insert a value greater than the previous ...
      ... one '}))
363   sup_k=input(' ');
364     end
365   end
366   else
367   end
368   sup(k)=sup_k;
369 end
370 end
371 clear sup_k
```

Translation of the phrase '*i* and *ii*' is simply '`sup_k>lowest_datum && sup_k<highest_datum && sup_k>0`'; as we have seen in the last paragraph of the previous chapter, condition *iv* in Matlab becomes '`isinf(sup_k)==0 && isnan(sup_k)==0 && isreal(sup_k)==1`' and condition *v* is represented by '`isscalar(sup_k)==1`'; it is obvious that the order constraint *iii* is translated into '`sup(k-1)<sup_k`'.

The for-loop from lines n° 315-316-317 to the end at line n° 370, wrapping all the other instructions, iterates both the waiting inputs for the elements of the partition, and the control on every one of them (followed by a message about the committed error), to build up correctly the partition.

At the end of each stage of the for-loop, the preference entered for a_j, initially configured as a constant `sup_k`, is redefined as a component ('`sup(1)=sup_k;`', '`sup(k)=sup_k;`'), depending on the index `k`, of the vector

`sup`. This trick is adopted to prevent Matlab from quiting the software if the enter button is wrongly pressed as an element of the partition.
Just inside, the conditional operator if-elseif at lines n° 318-342-369 executes the separation of the first point (corresponding to the first stage in the loop, '`k==1`', if-line 318) from the subsequent ones (corresponding to '`k>1`', at the elseif-line 342). Inside each branch there is a new, inner, if-elseif operator nested with a while-loop to verify whether the properties of the partition are satisfied: their positions are at lines n° 319-320-338-339-340 for '`k==1`', at lines n° 343-344-365-366-367 for the successive steps '`k>1`'. At n° 319-320 the commands if and while are followed by the logical complement of all the properties from *i* to *v*, except *iii*, connected to one another by the conjunction '`&&`'; at n° 343-344 there is the negation of all the logically conjoined properties. Such negation sentences at every step include all the possible errors on the inputs.

Chapter 6
Data analysis, part 1: data presentation, options' setup and principal outputs

From this point onwards, Italian firms' data at our disposal will be briefly presented (paragraphs 6.1 and 6.2) and it will be quickly illustrated how the parameters of the theories, of the methods and of the mobility measures, exploited in *MarkovInfer_MobInd.m*, have been configured (paragraph 6.3). We will look through the inventory of the most useful outputs which were built during the run (paragraph 6.4). Then, meaningful outcomes elaborated by the tool will be listed: some of them will be discussed for some Italian manufactures in the following chapters.
We processed, one at a time, 28 xlsx-type and single-sheet files, creating just as many Matlab workspaces. Any one of the workspaces contains the data matrices, all inputs selected by the user, some auxiliary objects supporting the completion of computations, and all the final outputs we are concerned with.

6.1 General presentation of the data

The 28 xlsx-type single datasheet were downloaded from the archive AIDA [1], for 5 Italian industries, as classified in the ATECO 2007 system, plus 2 groups of firms, not belonging to the ATECO taxonomy but published by AIDA itself. Only seven economic sectors have been looked into, due to lack of time. Incidentally, we remember that the ATECO 2007 taxonomy [4] is the classification of economic activities adopted by the Italian National Institute of Statistics (ISTAT) starting from January 1st , 2008; it was forerun by the classification ATECO 2002 [3] and is the national version of the Nace Rev.2, the European nomenclature sanctioned by the Regulation (EC) n° 1893/2006 of the European Parliament and of the Council, on 20 December 2006.
AIDA is the electronic archive pertaining to all Italian enterprises registered at the National Business Register of the Italian Chamber of Commerce, Industry, Agriculture and Artisanship (CCIAA) [5]. It has been our reference for the

setting of many practical tasks linked to the realisation of "*MarkovInfer_MobInd.m*". AIDA is developed by the company Bureau Van Dijk (see the home of the weblink [1]) in collaboration with InfoCamere [2], in-house company of the same Italian Chambers of Commerce, whose mission is providing information technology services. It is possible to find the monetary values (according to the principal international currencies) for variables of the complete Financial Balance Sheet and the complete Income Statement, as defined by the Italian Laws, for every company, together with other information like the fiscal code, the address and the geographical location. Whatever the login date, annual data for a decade are published; such a decade shifts one-year ahead, once a year at a conveyed date, on the beginning of March. Beside the two variants of the ATECO taxonomy (2002 and 2007), some not-European official classification systems for businesses, like the Nace-type ones, have been also implemented. Its internal browser allows for assembling any one of the innumerable possible matrices of data by matching the items of the user's preferred selection. Data matrices can be exported either in text format or in xls/xlsx-type spreadsheets. For these reasons, AIDA is extraordinarily exhaustive from a sampling perspective but cannot be employed in any inferential technique resting on time series, owing to the too much short lapse (only ten annual dates) of the shared detections. A multifunctional software, implementing the basics of statistical analysis to process archived information, is integrated but we did not avail of it, actually.

Within the overall panorama of Italian economic activities and according to ATECO 2007, the 5 industries are specific divisions of the manufacturing macro-sector (or manufacturing section, denoted by the preliminary letter 'C' in the 1st-level position, namely in the section position, of the detailed classification code). They were chosen for their high significance, both at the present and in an historical relevance, in any modern country:

- Food, cod. C10;
- Textiles, cod. C13;
- Chemicals and chemical products, cod. C20;
- Rubber and plastic products, cod. C22;
- Machinery and equipment, cod. C28;

(the numbers beside the names, just after the letter representing the section, are the digits associated to the ATECO division in the macro-sector, or 2nd-level digits, or section digits, of the code).

We resolved to inspect manufactures at the national scale, corresponding to the introductory two digits in the complete ATECO code, with the aim of having large original populations from where the maximal cores of incumbents could have been extracted, consisting of two thousand firms at least. Thus, it is presumed that any inferential procedure, grounded on the maximal sub-sample of agents actively operating for a decade, provides trustworthy results for the Italian national scenario.

The other two groups, staying outside the ATECO taxonomy (so, they have not any identification code number), are involved in innovative productions and are separately recorded in the Italian Register of Businesses:

- Small and Medium-sized Innovative Enterprises (Innovative SMEs);
- Innovative Start-ups.

Differently from the former five traditional industries, maximal sub-samples of permanent individuals from the last two groups in Italy are much less numerous, but innovative firms are interesting because they have special characteristics. However, all Small & Medium Innovative Enterprises, when including not-survived ones and later newcomers, is a very peculiar, well-determined population of low size, whose incumbents constitute a fairly-proportionate core. It is not correct to state the same for Innovative Start-Ups. The inferential analysis on them, compared to the much more crowed ATECO 2-digits sectors, might suggest intuitive preliminary considerations about the impact of sample size on statistical outcomes.

Irrespective of the industrial sector or group, the same 4 economic variables, fundamental in the financial statements of any firm, have been examined, *annually detected, and expressed in Euros*: two from the Income Statement, i.e. *Revenues from Sales and Services* (abbreviated in *Total Sales*) and *Total Production Monetary Value* (or, simply, *Total Production*), two others from the Financial Balance Sheet, i.e. *Total Assets* and *Total Payables* (or *Total Debts*). In the previous chapters the studied variable has been generically indicated as $X(t)$.

By severally processing the 7×4 single spreadsheets through our Matlab tool, we got 28 workspaces full of the outputs which derive from the deployed models, the tests and the mobility indices.

As explained in Chapter 4, our tool imports automatically all data from the original single spreadsheet, after receiving the input of its name, for every individual exhibiting a precise numerical value of the variable at least for one of the registered T observation dates. All data are saved in the matrix `All_Data`, from which the stationary core of all N incumbents from 2006 until 2015 (T=10) is acquired, saving it in the second matrix `incumbData`, sized N×10 (remember that in this thesis an incumbent is a statistical unit exhibiting a numerical entry for each of the 10 years).

The whole epoch 2006-2015, yearly registered in the AIDA archives at the time of our download, is maintained for each of the five 2-digit manufacturing divisions and for the group Innovative SMEs. In the case of the Innovative Start-ups group, instead, it is possible to obtain a stationary incumbent core uniquely for the shorter epoch from 2010 to 2015 (T = 6).

The number of incumbents N, and its proportion over the number of all agents in the original xlsx-type file, is primarily influenced by the sector, very weakly by the variable inside a fixed sector: they are placed in Table 6.1 for the possible combinations at our disposal. It is to be observed that N is copious, some thousands of firms anyway, for the five ATECO divisions (almost 4500 for Food Industry, even almost 8300 for Machinery and Equipment, between 2000 and 3300 for Rubber and Plastic Products, Textiles, Chemicals): a truthful inferential analysis is surely enabled on these sub-samples. In the area of the innovative entrepreneurship, outside ATECO classification, the ecosystem of Innovative Small and Medium Enterprises is, by its very nature, far less crowded, in practice 300 firms, than the previous categories: it is a very well specified and low-sized sub-population, as we have anticipated above in the paragraph. But the percentage of its 115 incumbents over the total (~38%) is comparable to the cases of Food, Textiles, Rubber and Plastics (providing percentages of residing units from ~30% to ~50%). Therefore, we are convinced that any inferential result about the latter six can be considered reasonable in an equal extent, because they are fairly representative of respective sources. The situation of Innovative Start-ups is opposite, because in

Italy they have been strongly encouraged in the most recent years, so in our spreadsheets 4300 start-ups are reported. Unfortunately, many of them last only for a short time as autonomous firms, (a few years after their birth, very often they will fail or will be acquired by a better-established company; and it happens whatever the country). Hence, we can extract a dozen of incumbents, which is a very exiguous proportion of the total, from 2010 to 2015 and not before. It is undoubtedly likely that such an incumbent panel misrepresents any stochastic law underlying the intrinsic dynamics in an innovative-startup ecosystem; nevertheless, it might be assumed as a very rough benchmark to compare data analysis's outcomes, under the condition of an extremely poor statistics, with the other much more populated cases.

Industry or Group (ATECO 2007 code) Economic Variable *X*(*t*) (in Euro)	Food & Beverages (C10)	Textiles (C13)	Chemicals (C20)	Rubbers & Plastics (C22)	Machinery & Equipment (C28)	Inno SMEs	Inno Start-ups
Total Sales Monetary Value-number of All firms	13758	5736	4026	7116	16914	301	4307
Tot. Sales Monetary Value- numb. of Incumbents *N*	4486	2728	2045	3286	8243	115	12
Total Assets- number of All firms	13758	5736	4026	7116	16914	301	4307
Total Assets- number of Incumbents *N*	4497	2729	2048	3291	8251	115	14
Total Debts- number of All firms	13758	5736	4026	7116	16914	301	4307
Total Debts- number of Incumbents *N*	4497	2729	2047	3291	8251	115	14
Tot. Product. Monetary Val.- numb. of All firms	13758	6184	4207	7116	18107	302	4310
Tot. Product. Monetary Val.- numb. of Incumb. *N*	4493	2747	2050	3290	8275	115	12

Table 6.1: the number of all agents and the corresponding number of extracted incumbents, for each economic activity and every variable

6.2 A slightly more in-depth comment about AIDA samples

We imagine every industry at the national scale (identified by the beginning letter and the next two digits in the ATECO 2007 code) as an autonomous ecosystem (at least, we believe to have chosen five Italian ATECO divisions not strongly correlated among one another, every one of which is connotated by its own peculiarities), whose historical events (any factor influencing the industry and giving rise to its changes) were occurring from 2006 up to 2015. The ecosystem is inhabited by its own population (of firms) which is not constant with time because, after the starting date of the decade, some old

inhabitants go out from the ecosystem (in other words, some firms may leave the AIDA datasheet of the industry) while some new inhabitants are going to enter it (corresponding to new firms emerging in the same datasheet). It has been remarked in paragraph 2 of the Chapter 4, treating the automatic import, that any sequence of appearances and disappearances from official archives is a more generic and '*tainted*' phenomenon, including mergers and acquisitions along with notification flaws from economic subjects in a difficult situation, than the genuine dynamics of entries and departures respect to the market. In order to overtake such a hindrance, from the dataset of each ATECO national division the core of enduring inhabitants, on the whole age 2006-2015, has been extracted by *MarkovInfer_MobInd.m* (they are the spreadsheet's incumbent enterprises).

Now, to comprehend both the reasons in the previous paragraph and the comments in Chapter 7, it needs to remind the problem of sample-size impact on inferential procedures and hypothesis tests. Within the community of statisticians, it recurs the idea that small size, large size or huge size of a sample are concepts with an absolute sense; instead, in our opinion, it is fundamental to interpret any size as a relative concept, to be gauged to the size of the population to be investigated. Moreover, such a problem is linked with the issue of the definition of the population itself and its correlation with some other ones. There exists the concrete risk that a sample of some ten thousands of units is not able to represent a wide population of millions and millions of individuals, owing to the small ratio and to the high degree of heterogeneity of the population itself; whereas an intriguing but little and very distinguished sub-population of some hundreds of units may be represented more exhaustively than a much greater one, from a statistical point of view, through 100 or 200 of its agents (which are a relatively high proportion), homogeneous to their original low-sized source.

We are sure that, for each industry in AIDA, the incumbents' sub-sample in practice coincides with the entire real sub-population of permanent businesses in the market (*equivalently, the former is a highly-representative statistical panel*) *since, if some firms residing in the market during the decade were not registered into the AIDA electronic sheet, they are so few not to affect significantly any inferential estimations*. We are convinced the same thought is right for the Innovative SMEs, which is a well-determined group outside the

ATECO classification, albeit it is a low-sized sub-population: it is likely for us that almost all the small and medium innovative companies, acting in the Italian market from 2006 to 2015, are registered in AIDA. Consequently, *the core of 115 residential Innovative SMEs, versus a varying total sample of 301 not-permanent companies on the decade, must be considered a good statistical representation of a very specific* (*though small*) *population, rather than an absolutely scarce panel. Indeed, inside the InnoSMEs group the size proportion between incumbents and the complete sample is very similar to the analogous proportion of the ATECO national sectors* (*from 30% to 50%*). On the contrary and without any doubt, the sole 12-14 Innovative Start-Up, enduring in AIDA between 2010 and 2015, in comparison with the total population of approximately 4300 Innovative Start-Up appeared in the same age, are a too scant panel to be an acceptable statistical representation: that is why elaborations about them are simply an instance of misleading inference and a rough counterproof.

6.3 Configuration of the 'free parameters'

In the first and the second paragraph of Chapter 3 we introduced the main characteristics of the device *MarkovInfer_MobInd.m*, disclosing in paragraph 3.1 the sequence of stages in which its complete run is organised; the first paragraph of Chapter 4 begins with the catalogue of settings to be adjusted by the user, according to his preferences, into the prompt of the command line waiting for an input, in order to automatically compute, and save into the Matlab workspace, the set of outputs of our interest. Here we are going to specify values and shapes we decided to assign to the settings contributing to the session of statistical examinations commented in this thesis.

Deciles have been chosen, entering '10' at the waiting prompt as input for the scalar `qntl_degree`, to enlighten the distribution of the data for the four variables and for all economic activities, except for Innovative Start-Ups. For the latter, we faced a double problem: it happens not only their core of incumbents is extremely exiguous. Generally, special balance-sheet rules and protections are initially granted, since start-ups arise in an unfavourable competitive scenario and try to survive in challenging conditions. Thence, it may also happen some items inside the Financial Balance-Sheet and Income Statement of many start-ups are zero for some years; consequently, some

deciles, preceding the sample mean, may coincide with zero itself. Owing to our tool needs an ordered not-trivial sequence of positive quantiles to operate, for Innovative Start-Ups we adopted quintiles, entering '5' at the prompt for `qntl_degree`, *ceteris paribus*. However, deciles can be aggregated to obtain some rougher distributions for the same panel. Obviously, from a single spreadsheet a set of deciles, for every one of the T observation dates, is estimated; those sets are arranged as rows into a $T\times10$ matrix (it is a $T\times5$ matrix for the start-ups), named `Quantiles` (see the third paragraph in Chapter 4), where a particular column corresponds to one, and only one, position in the deciles' ordering; the dates vary from 1 to T when descending from the top to the bottom inside any column.

Deciles (or quintiles) have been opted also to be extremes $\{a_1, a_2, \ldots, a_{m-1}\}$ delimiting the intervals $\{ [a_j, a_{j+1}), j = 0, 1, 2, \ldots, m-1\}$, in their turn functioning as m stationary Markov states, into which the economic variable's domain would have been divided. Hence, we entered the value '10' (or '5') also for the parameter `classes_num` representing the number of Markov strata m in our program. But since there is a set of quantiles for each observation date recorded in a spreadsheet, we always inserted at the subsequent interactive step, as inner entries of the vector `partition` (see Chapter 5, first paragraph), the ordered set of deciles (or quintiles) containing at the tenth (or the fifth) position the greatest variable's value, $\max_{\forall t,\forall l} X_l(t)$, respect to the whole datasheet, whatever the observation. The program is able to track down and display such maximum among all the numerical cells inside the xlsx-type file, storing it into the workspace as a scalar called `highest_datum` (Chapter 4, paragraph 3). All inferential estimations and statistical tests dealing with Markov chains and individuals' residence durations, as well as mobility calculations according to the diverse indices, are carried out on the basis of the 10 (5 for startups) classes determined by the set of quantiles including the maximum. Due to the sure absence of any agent beyond the absolute maximum, that is inside the half-line $[a_m, +\infty)$, the upmost class applied by *MarkovInfer_MobInd.m* is $[a_{m-1}, +\infty)$, in practice equivalent to $[a_{m-1}, a_m)$.

The next parameters to be sequentially entered are the weights $\{\omega_i, i = 1, 2, ..., m=10 \text{ or } 5\}$ of the possible transitions' starting states, which fall within the definitions of the Directional Mobility Index (Ferretti and Ganugi (2013)) and the Mobility-as-Movement Index (Alcalde-Unzu, Ezcurra, Pascual (2006)), while the other indices does not depend on departure conditions. Our software was written so that both the indices share the same set of $\{\omega_i\}$, which are assembled into the vector `w_k`, to make simpler the comparison between them. Moreover, the uniform distribution when $m=10$ has been fixed to allow for sensible comparisons among all the mobility measures based on transition matrix we are interested in. In fact, apart from Directional Mobility and Mobility-as-Movement, definitions of the other indices emphasise the sole dependence on the transition probabilities without contemplating any impact of the starting positions on the mobility concept. Consequently, the helpful manner of minimizing the importance of departure conditions in the Ferretti-Ganugi Index and in the Alcalde-Unzu one is the choice $\{\omega_i=1/10 \ \forall i = 1, 2, ..., m\}$.

Another factor, defining a further reliance, in the Ferretti-Ganugi and the Alcalde-Unzu Indices, that is absent from the rest of transition-matrix-based measures, is $v(|\ i - j\ |)$ the modulating function which weights the contribution of transition breadth. It is also shared by the two mobilities above and is represented in the software by the bi-dimensional array `V_k_h`. We decided the modulating function to be shared in the device, too. We also refused to diminish the complexity of the two indices and to make them akin to someone of the other seven: so, the trivial choice $v(|\ i - j\ |)\equiv 1$ was rejected. Besides, a stronger law than linearity $v(|\ i - j\ |) \equiv |\ i - j\ |$ has been preferred, with the aim of highlighting the influence, on the mathematical description of mobility, of jumps' amplitude respect to its lack; nevertheless $v(|\ i - j\ |) \equiv |\ i - j\ |^2$, a quadratic function, would have made jumps' amplitude to prevail over related transition probabilities.

Finally, the compromise has been $v(|\ i - j\ |) \equiv |\ i - j\ |^{10/6}$.

The last inputs to be supplied are the power order α, applied element-by-element to the absolute value of the difference between the transition matrix and the identity, in the Alcalde-Unzu Index; and the power order β in the Exponentiated (or Amplified, or Generalized) Determinant Index, conceived in

Shorrocks (1978) to be period-invariant. We have fixed for those parameters two values which are greater than 1 but also clearly smaller than 2, avoiding the banal choice $\alpha = \beta = 1$: precisely $\alpha = 1{,}3$ and $\beta = 1{,}5$.

6.4 Compendium of meaningful outputs, built up from the data

Recall that the number of incumbents N varies according to both the industry and the variable $X(t)$. The number of detection dates is the same, $T = 10$, for all sectors both in the spreadsheets and in the workspaces, except for the Inno-startup group, for which it reduces to $T = 6$ in the four workspaces after the incumbents' pulling out. The number of Markov states, marked by deciles, is $m = 10$, but Innovative Start-ups are different again because the states are marked by quintiles, so $m = 5$.

For every one of the 28 workspaces elaborated from a single xlsx-type datasheet, an overview about the most intriguing final outputs for a statistical investigation is listed just below (the most of intermediate objects is omitted).

- The matrix `All_Data` consisting of observations for the totality of registered firms, distributed on T columns (one row for one firm; for some firms, the numerical datum may be unavailable at certain dates), whose number is represented by the first component of the two-element vector `dim_All_Data`.
- The matrix `incumb_Data`, sized $N \times T$, constituted of those N firms showing a numerical datum of the variable for all the dates; the number of rows of such matrix is also saved into the scalar `num_of_incumb`, whose share respect to the total number of firms is the scalar `share_of_incumb`.
- Two vectors, sized $1 \times T$, `MinValues` and `MaxValues`, respectively collecting the smallest observation and the greatest one for each date in the spreadsheet;
- three $1 \times T$ vectors named `Means`, `Medians` and `Modes` gathering, for all dates, the means, the medians and the modes of the data in the spreadsheet;
- the matrix `Quantiles`, sized $T \times m$, composed of the rows of data deciles for every date;
- the $(m+1)$-component vector '`partition=[0, Quantiles(t highest_datum,1), Quantiles(t highest_datum,2), Quantiles(t highest_datum,3),.., Quantiles`

(t highest_datum,8), Quantiles(t highest_datum,9), Inf]', where t highest_datum is the date of the maximal datum inside the entire incumbent panel (if the classes are determined by the quintile set, the last element before infinity is Quantiles(t highest_datum,4)).

- The *m*×*m* matrix p_stat_k_h constituted of the maximum likelihood estimations of the first-order, *stationary* transition probabilities from any *k*-th state to the *h*-th one: p_stat_k_h is the transition matrix of the 1st-order *stationary* Markov model based on the data.
- The *T*−1 matrices, sized *m*×*m*, of the maximum likelihood estimations of the first-order, *not-stationary* transition probabilities from any *k*-th state to the *h*-th one; every matrix is evaluated between the starting time t_start and the successive t_start+1 (everyone is *one-period* or *one-step* matrix); actually, such matrices are 'laid one upon another' inside a 3-dimensional array, named p_nonstat_k_h_tstart and depending on the three distinct indices (k,h,t_start), where t_start works as the third dimension and assumes all integer values from 1 to *T*−1; in other words, p_nonstat_k_h_tstart is the ensemble, ordered respect to the time t_start, of all one-period transition matrices from the 1st-order, *not-stationary* Markov model inferred from the data.
- The *T*−1 multiperiod transition matrices, sized *m*×*m* everyone, relative to the 1st-order *stationary* Markov model and calculated from the first date in the spreadsheet, t_start = 1, to each of the following dates t_ult+1= 2, 3,.., *T*: each matrix is the power of order t_ult of the one-period matrix p_stat_k_h; power_p_stat_k_h_tult is the 3-dimensional array, depending on the indices (k,h,t_ult), where the matrices are put together and ordered respect to the time t_ult.
- Another set of *T*−1 multiperiod transition matrices, whose size is *m*×*m*, relative to the 1st-order *not-stationary* Markov model: each of them is the product of all the one-period matrices in p_nonstat_k_h_tstart from the matrix corresponding to t_start=1 until the one corresponding to t_start = t_ult, which varies from 1 to *T*−1; aggr_p_nonstat_k_h_tult is the 3-dimensional array, depending on

the indices `(k,h,t_ult)`, where the multiperiod matrices are arranged in order according to the time `t_ult`.

- The scalar `df_timedepend_tot` representing the number of the system's degrees of freedom (the system is the panel of incumbents) for the null hypothesis that the 1st-order Markov chain is stationary, against the alternative that the 1st-order chain is not-stationary; `df_timedepend_tot` is used both for the chi-squared test and for the maximum likelihood ratio criterion;
- the scalar `df_1stVS2nd_tot` representing the number of the system's degrees of freedom for the null hypothesis that the stationary Markov chain is of order 1, against the alternative that the stationary chain is of order 2; `df_1stVS2nd_tot` is adopted both for the chi-squared test and for the maximum likelihood ratio criterion.
- The scalars `chisq_timedepend_oss` and `MLRcrit_timedepend_tot` are the statistics, evaluated for the current data, relative to the same null hypothesis that the 1st-order Markov chain is stationary, which admit the chi-squared distribution as an asymptotic probability distribution, respectively for the chi-squared test and for the maximum likelihood ratio criterion;
- the scalars `chisq_1stVS2nd_oss` and `MLRcrit_1stVS2nd_tot` are the statistics, observed for the data, relative to the same null hypothesis that the stationary Markov chain is of order 1 but is not of order 2, which admit the chi-squared distribution as an asymptotic probability distribution, respectively for the chi-squared test and for the maximum likelihood ratio criterion.
- Referring to the null hypothesis of 1st-order stationarity of the chain, `p_X2_timedepend_tot` and `p_MLR_timedepend_tot` are the probabilities, respectively coming from the chi-squared test and the maximum likelihood ratio criterion, that such null is true, i.e. the probabilities that the chi-squared random variable, distributed with a number of degrees of freedom equal to `df_timedepend_tot`, is respectively greater than the associated empirical statistics `chisq_timedepend_oss` and `MLRcrit_timedepend_tot`; in our device it has been employed the threshold of 5%.

- Concerning to the other null hypothesis of 1st-order stationarity against 2nd-oeder stationarity of the Markov model underlying the dynamic of the incumbent panel, `p_X2_1stVS2nd_tot` and `MLRcrit_1stVS2nd_tot` are the probabilities, respectively from the chi-squared test and the maximum likelihood ratio criterion, that second null is true, in other words the probabilities that the chi-squared random variable, distributed with a number of degrees of freedom equal to `df_1stVS2nd_tot`, is greater than the corresponding empirical statistics `chisq_1stVS2nd_oss` and `MLRcrit_1stVS2nd_tot`; even in this case, it has been employed the threshold of 5%.
- The 1×*m* vector `mean_permtime_M1stat_k` of the mean permanence times spent by an incumbent in each *k*-th stratum, along with the 1×*m* vector `sigma_mean_permtime_M1stat_k` of the standard deviations matched one-by-one to them, estimated from the data through the stationary, 1st-order Markov chain as it is suggested in equations (5) and (6) of Prais (1955) as well as in its Appendix B at pages 65-66;
- The 1×*m* vector `mean_permtime_M1nonstat_k` is proposed as an alternative formulation for the mean residence durations of an incumbent in each *k*-th class, along with the 1×*m* vector `sigma_mean_permtime_M1nonstat_k` of their matching standard deviations, because they are estimated through the not-stationary, 1st-order Markov chain by adapting within reason the stationary 1st-order model. More precisely, the series in Appendix B in Prais (1955) were broken off to obtain a finite sum of as many terms as the number of periods registered in the datasheet, while the different powers of the stationary transition matrix were being replaced with products of not-stationary transition matrices associated to consecutive observation periods.
- The third 1×*m* vector `mean_permtime_observed_k` does not derive from any Markov approach, but it comes from a simple mechanistic computation of the empirical mean time spent by an individual of the fixed panel to remain in the same state between two consecutive dates; sample standard deviations of all components of such vector are arranged in its companion `sigma_mean_permtime_observed_k`; they both depend on the only index '`k`', like the previous ones. This mechanistic calculus is

supported by an essential multi-dimensional array called `Agent_k_h_tstart`.

- `Agent_k_h_tstart` is a 4-dimensional array, depending on the indices (`Ag_id,k,h,t_start`), whose elements typifies the transition of every single statistical unit between `t_start` and `t_start+1`, namely which is worth 1 if the agent, identified by a specific value of the parameter `Ag_id`, occupied the state '`k`' at the date `t_start` and has been found at state '`h`' at the time `t_start+1`; thus, if the agent '`j`' remains in the class '`h`' from `t_start` till `t_start+1`, it holds '`Agent_k_h_tstart (j,h,h,t_start)=1`'.
- The $1 \times m$ vector `w_k` whose entries are the weights $\{\omega_i, i = 1, 2, ..., m\}$;
- two $m \times m$ matrices `V_k_h` and `Vs_k_h`, respectively implementing in *MarkovInfer_MobInd.m* the modulating functions $v(|\, i - j \,|)$ and $\text{sign}(i - j) \cdot v(|\, i - j \,|)$ appearing inside the Alcalde-Unzu Index and the Directional Index;
- the nine scalars representing the nine, transition-matrix-based, mobility indices measured from the incumbents' data, according to the 1st order stationary Markov chain, that is they all are calculated on the transition matrix `p_stat_k_h` and do not depend on any time parameter: they are arranged into the column vector `stat_Mobilities`.
- Nine row vectors, $1 \times (T-1)$ everyone, are the nine mobility indices according to the 1st order not-stationary Markov model, for the imported incumbents' panel: each row is a particular mobility measure calculated between any pair of dates (`t_start`, `t_start+1`) registered in the spreadsheet and each component of a row is a certain type of mobility based on the transition matrix `p_nonstat_k_h_tstart` (so, the entries of every row are one-period, or one-step, mobilities). Such vectors are collected one upon another in the bi-dimensional array `nonstat_Mobilities`.
- A further set of nine, $1 \times (T-1)$ vectors is the set of the aggregated, or compound, forms of the nine mobility indices in the paradigm of the 1st-order Markov stationarity for the usual incumbents: each row refers to a peculiar mobility definition implemented for the powers of the stationary matrix `p_stat_k_h` composing the array `power_p_stat_k_h_tult`.

These powers correspond one-by-one to the composition of transition phenomena from `t_start`=1 until every posterior date `t_ult+1`=2, 3, …, *T* (it is the reason why the components of the rows are compound, or aggregated, stationary mobilities). The vectors are gathered inside the bi-dimensional array `Aggr_stat_Mobilities`.

- The set of nine, 1×(*T*−1) vectors of the aggregated (compound) variants of the nine mobility indices in the framework of the 1st-order, Markov, not-stationary model: each row refers to a peculiar mobility definition computed for the products of the not-stationary matrices `p_nonstat_k_h_tstart` forming the array `aggr_p_nonstat_k_h_tult`. Since these products correspond one-by-one to the composition of transitions from `t_start`=1 until every future date `t_ult+1`={2, 3, …, *T*}, the components of the rows are compound, or aggregated, not-stationary mobilities. Such vectors are placed inside the bi-dimensional array `Aggr_nonstat_Mobilities`.

Chapter 7
Data analysis, part 2: 1st-order stationarity *vs.* not-stationary 1st-order and 2nd-order stationarity

7.1 Trust or not trust Markov time-homogeneous 1st-order in Economics? An essential summary

It is stressed in paragraphs 1.1 and 1.2 of Chapter 1 that, during the 1950s and the 1960s, the stationary 1st-order Markov chain became an intensely investigated stochastic model to describe the dynamics of economic quantities; rather, it became the stochastic model '*par excellence*' in economic sciences. In this paragraph we are going to revisit the outcomes, in its regards, of statistical verifications conducted by our bibliographic references. *We want to emphasise* in those studies the same features involved in our analysis (see the previous Chapter 6): *type of industry* and *variable*; *sample size*, *country* and *epoch* the data refer to; whether it is *available a fixed sample* of firms, each of them expressing a datum for every observation date, *or* whether individuals' entry-and-exit phenomena, compared to the initial sample' composition, passing the time, are included; and last, but not least, the *final* '*scores*' turned out *by reliability tests*, when provided.

Feedbacks appeared initially positive. In Prais (1955) for the first time it is used as a theoretical background to study the from-fathers-to-sons mobility among 7 social cohorts, as well as the mean permanence times inside them, for a sample of 3500 males resident in England and Wales and interviewed in 1949. In Adelman (1958) such chain, integrated by an input-output reservoir, is applied to elaborate total assets' data dealing with U.S. steel industry from 1929 to 1939 and from 1946 until 1956, forecasting a long-run distribution, based on the idea of dynamic equilibrium, which is judged fair-minded if compared to the context and trends implicit in the data. But these two works do not provide any statistical test about any Markov feature, differently from the

inferential study for stationary 1st-order chains having a stochastic mechanism to account for individuals which enter the sample or leave it, after the starting date, by Duncan and Lin (1972). Here the Markov classes are determined through the ratios of bank's loans to farms over the net total loans for the Ninth Federal Reserve District banks, from 1954 to 1969, for a sample whose size varies until the maximum of 528 banks. Many stochastic processes for entries into any one of the active classes are contemplated and an additional absorbing state for all exits, without re-entry opportunities, is put next to them. Then, maximum likelihood estimations and likelihood ratio criteria, originally proposed in Anderson and Goodman (1957) for multinomial parameters and a fixed sample, are adapted (pages 762 and 766) to such new extended Markov paradigm. The authors declare that "highly significant probabilities" (page 765), found out from maximum likelihood ratio tests, absolutely validate the hypothesis of a 1st-order stationary Markov system against the alternative of 2nd-order and pointwise time-dependent transition probabilities. Precisely, under the null that 1st-order transition probabilities are constant in time, the likelihood ratio statistics λ is asymptotically distributed likewise a χ^2 random variable with 300 degrees of freedom and the λ value from the data is the significant 368,8. Conversely, the joint test of stationarity and 1st-order against not-stationarity and 2nd-order yields an observed λ value of 787,3 in the case of 1475 degrees of freedom, showing the alternative to be very little relevant. Moreover, the Pearson's goodness-of-fit tests, confronting predicted frequencies distributions with the observed ones in 1968 and 1969, is in favour of the former model, embodying stochastic inputs and outputs, "the values of the [associated] statistics [being] all smaller than the critical table value $\chi^2_{95\%}(4) = 9{,}49$" (p. 765).

Anyway, in paragraphs 1.1 and 1.2 we have also focused on the fact that these testing results are not confirmed at all by another part of the scientific literature, from the late 1960s, decade after decade, until today.

In Hallberg (1969) it is discovered, right by the maximum ratio criterion for time dependence in Anderson and Goodman (1957), that the assumption of transition probabilities to be constant with time must be rejected for the annual sales volume (in gallons) of the plants belonging to a not-fixed sample and manufacturing frozen-milk products from 1943 to 1963 in Pennsylvania. Markov first-order stationarity leads to erroneous predictions on 1964 and on

1965, even if a simple entry-exit repository, occupied by plants without any production during a given year, is placed alongside the active four classes of positive sales (the greatest size reached by the sample is 884 plants). Besides, the computed value of the statistics $-2\log_e\lambda$ (developed in equation (6) at page 291), over the 20-year period, is $\chi^2_{data} = 749{,}14$, too much greater than the number of 360 degrees of freedom in order not to refuse the first-order stationarity.

The analysis in Shorrocks (1976) casts doubts on the conjecture that mobility is a time-independent first-order Markov process, as a rule, respect to incomes, preferring a second-order chain based on a reinterpretation of the postulates. It relies on evidences for the observed transition rates from a British fixed sample of 800 coetaneous male employees, whose annual incomes are known only for the years 1963, 1966 and 1970. Maximum-likelihood stationary estimates of both the 1st-order transition probabilities and of some of the 2nd-order conditional probabilities over a triplet of states were computed. In the former model all the possible jumps are treated; in the latter it is exclusively inspected the mobility restricted to a single-step movement in every possible direction, for every one of two couples of dates. The results of performed likelihood-ratio tests, collected in Table 1 at page 575, show that the first-order process performs badly and should be discarded for all seven classes examined, because the logarithms of all likelihood ratios are greater than the critical values 39 and 44 of a χ^2 random variable with 60 degrees of freedom, corresponding to the respective 5% and 1% significance levels.

Bickenbach and Bode (2003), through statistical inference on time series relying on the hypothesis of strong stationarity of the income stochastic process (and, broadly speaking, of all its functions), starting from a 'regional' geographical scale, have contested the idea that income pacing is well approximated by the 1st-order time-independent Markov framework. In their in-depth and articulated work, those authors have applied chi-squared tests and maximum likelihood ratio criteria to the Markov property, to the spatial independence, and to homogeneity over time and space for first-order transition probabilities focusing on relative regional per-capita incomes across the 48 contiguous U.S.A. states, from 1929 to 2000. So, 3408 aggregated observations are obtained from the available, fairly long, time series. All tests, inspired by the Anderson and Goodman's inferential researches, implement an

important correction: whenever any estimated parameter, located as a divisor inside the statistical ratio, is worth zero, then the related Markov stratum is excluded from computations of both the statistics and the number of degrees of freedom, whatever the null statement. Incidentally, such a correction is employed in *MarkovInfer_MobInd.m*, too. To avoid any 'poor statistics situation' in the data analysis, intervals longer than one year have been adopted, even if the time in the data is annually marked. It should be noted that the failure has been attributed not only to the restrictive nature of the Markov postulates, but also to a varity reasons linked to the complexity of interactions among states inside the U.S. federation as well as between U.S.A. and the international scenario. The consequence is a lack of homogeneity respect to time and space. Indeed, two structural breaks happened in the registered historical epoch, the major just after the World War II, significantly conditioning the evolution of the income distribution; a later minor one in the late 1990s. A different development is exhibited by some groups of states respect to other ones; development type changes if a state is surrounded by rich neighbours rather than by poor neighbours. Degrees of freedom and totals of both criteria' statistical ratios are painstakingly arranged in their Table 2 (page 376), while contributions from each Markov stratum to the same two statistics are in their Table 3 (page 377), for the null hypothesis of time homogeneity versus time dependence; for the other null hypothesis of history independence (Markov first-order versus Markov second-order), all the test parameters are organised in Table 5 (page 379). For both such questions, the most frequently ascertained fact is the Pearson's ratio and the maximum likelihood ratio to be higher than the number of degrees of freedom, denoting the two null hypotheses as incompatible with U.S. income patterns, at least when a great amount of observations is available.

7.2 General and peculiar considerations about hypothesis tests in *MarkovInfer_MobInd.m*

To check up the fundamental premises underlying the Markov chain approach for modelling the dynamics of a system, researchers can take advantage of a variety of methods. We chose to compare maximum likelihood (ML) estimations of transition probabilities under two specific couples of a null

hypothesis and its alternative by the maximum likelihood ratio criterion (MLRC) and the chi-squared test, according to the lessons in Anderson and Goodman (1957), Goodman (1958), Billingsley (1961). The two couples of the form [null hypothesis-alternative one], we have adopted in the Matlab tool, are:

i. time-homogeneous 1st-order chain against a 1st-order not-stationary one (*Query a* or *Antinomy a*);

ii. time-homogeneous 1st-order chain against a time homogeneous 2nd-order one (*Query b* or *Antinomy b*).

In Anderson and Goodman (1957) it is proved that the chi-squared test statistics is asymptotically equivalent to the MLRC statistics when matching Markov chain's characteristics, but they may remarkably differ in cases of "*poor asymptotics*": such a situation does not occur for data of the 5 'classical' manufactures at our disposal, pertaining only to our data of Innovative SMEs and Innovative Start-ups. Anyway, both the MLRC and the chi-squared statistics for ML estimators will be calculated by the device and discussed.

It is better to clarify that reliability of estimations of the transition matrix, in conformity with a certain theoretical framework, depends at least on two determinants, i.e. the real stochastic process generating the data and the number of observations dealing with the transition. It is needed that the real process, from which the sample is drawn, satisfies all the restrictions imposed by the underlying theory which is, here, the Markov one. Otherwise, the ML Markov estimations neither of the not time-homogeneous conditional probability $\hat{p}_{ij}(t)$ from any time t until the next t+1 nor, least of all, of the time-homogeneous probability $\hat{p}_{ij}$, will not be an authentic estimator, for every starting state i, of the actual transition probability distribution from i itself to any destination state j. The necessity to obey the Markov restrictions makes it not for granted at all the system under study to be of Markov type. Furthermore, the ML estimation must be based on a large enough number of occupiers, $n_i(t)$, of the initial state at the date t (and consequently on a fairly large number of agents n_i^* starting from i at any date) to allow for relying on the asymptotic properties of the estimators; or else the accuracy of estimates will be poor. Nevertheless, in practice it happens that the more numerous the sample drawn from the whole system, the more unlikely it is that a theory, selected *a priori*, turns out to be completely trustworthy respect to the data. Thus, any estimation procedure seems basically to be a compromise between improving the accuracy of

estimates and increasing the probability of violating the assumptions underlying the preselected paradigm which should describe the examined phenomena. This deadlock can be overcome comprehending that, in principle, it is preferable to estimate the probabilities using as many observations as possible. Obviously, taking such a decision, it would be advisable evaluating a range of theoretical models (not a unique model) and confronting them with one another to choose the best, but it might be a hard and long-lasting challenge.

Datasets from AIDA, derived from official registers and archives of the Italian Chambers of Commerce about Italian industries and firms, are befitting to avoid drawbacks mentioned above. For every observation date, every industry and geographical area, the maximal sample reported in AIDA is very similar to the effective population (remember that some firms, experiencing problems or unexpected circumstances, sometimes may not communicate their data to InfoCamere). In case of the greatest manufactures, after extracting all the units showing a datum at each date during the published ten-year epoch, the subsamples are still big enough in prospect of any right inferential analysis, having some thousands of firms. Hence, if it was disclosed by our hypothesis tests the first-order stationary chain to lose versus the alternatives, it would not be the fault of the sample size.

Statistical economists, interested in empirical researches, agree on the intuition that the shorter the time intervals constituting the transition periods, the higher the probability the system dynamics to depend on time, namely the higher the probability of violating the essential Markov property; ergo the first-order stationary chain becomes a bad representation of the system. We are convinced the one-year periods, always subdividing the decade published by AIDA, are not so short as to accentuate by themselves contingent deflections of the system from the first-order time-homogeneous chain. Hence, if it was revealed by our tests such Markov chain to lose versus the not stationary first-order and the stationary second-order, it would not be the fault of the periodization, but it is something of intrinsic.

Besides, in AIDA time series are so limited and the samples are so extensive that it is licit to believe that sample-size effects tend to favourably prevail over possible ‘periodization distortions’, when executing any inferential procedure on its data. In this regard, the sole considerable exception are structural breaks

because the longer the inspected historical era, the higher the risk of structural breaks, whose existence implies a not negligible heterogeneity respect to time. Indeed, it is already noted in Fingleton (1997) that the Markov chain approach is suited to absorb the impact of an irregular stream of small shocks sometimes affecting economies; big rare shocks, instead, are not consistent with time invariance of transition probabilities. Concerning the Italian data from 2006 until 2015, an extended break into the national economic structure has occurred in concomitance with the delayed propagation of the Great Recession to Italy, after the crisis from U.S.A. (where it had begun in 2007) spread to Europe. Owing to the very modest financialization of the national economy, the Great Recession has developed in Italy starting from the 2009 (in 2011 the country has been put *de facto* under the European Union's supervision) until the end of 2015, when the last 'dramatic' decrease of its G.D.P. has been officially measured. It has appeared as an alternation of annual falls and annual weak recoveries of the G.D.P., which implies the general trend of a noticeable decline. On the one hand such great shock from 2009 to 2015 might influence our reliability investigations to the detriment of 1st-order Markov stationarity; on the other, if the two alternatives will predominate in virtue of very near probabilities to 100%, it is very likely that inadequacy of the 1st-order stationary chain is mainly intrinsic instead of partially determined by the break of the Recession. At any rate, there was not enough time to separately analyse the subperiods 2006-2008 and 2009-2015, although it would be of huge interest.

It is helpful another preliminary reflection. If the Markov approach of the first order is refused against the second one, inferential tests could be broadened to higher transition orders introducing further dimensions associated to times t–3, t–4, and so on. However, it is needed to pay attention that, for a given dataset, when such a sequence of backward time-steps becomes more and more numerous, the number of parameters to be estimated exponentially increases, but the number of disposable observations linearly decreases. Consequently, the credibility of estimates and the power of the test rapidly decreases. For this reason, it is advisable to set an *a priori* limit up to which to search the most fitting Markov order, as suggested in Tan and Yilmaz (2002): we decided to stop at the second order.

Finally, it is to be highlighted that, as far as we know, it does not exist yet any statistical test to directly contrast the two winning alternatives with each other, namely the 2nd-order time-homogeneous chain with the 1st-order not-stationary chain. Hence, it was necessary to choose one of them to carry on compiling the programme and concluding the present work with the topic of mobility indices: we opted for the not-stationary 1st-order model, since transition-matrix based mobility measures, established upon it, are more straightforward and well-proven.

7.3 Results of hypothesis tests via *MarkovInfer_MobInd.m*

By processing datasheets concerning the samples of residing Italian companies (or incumbents) during the period 2006-2015, whose outputs are arranged in Table 7.1 and Table 7.2, some surprising regularities emerge from the inferential verifications about the Markov features implemented in *MarkovInfer_MobInd.m*, relative to the four balance-sheet proxies for all the five ATECO 2007 manufactures (see the five central columns in the Tables). It is to be repeated that Innovative Start-Ups and Innovative Small & Medium Enterprises are businesses' groups contemplated inside the archive AIDA, but they do not fall into the ATECO classification. These recurrent characteristics involve both the queries we have considered dealing with the chain, which admit the feature of 1st-order time-homogeneity as a null hypothesis in contrast with the alternatives of 1st-order not-stationary chain (*Query a*, Table 7.1) as well as 2nd-order stationary chain (*Query b*, Table 7.2), for the chi-squared test and the equivalent maximum likelihood ratio criterion (MLRC). Let us remember that, for every variable in every industry, the Markov state space is created by segmenting its maximal domain by the distribution's deciles at that date. Since resoundingly diverse trends arise from outcomes of the same statistical examinations on those two separate innovative groups, when processing their datasheets by the same program, we are sure that regularities in the 20 ATECO fixed panels are not determined by any systematic error in our Matlab code. But they are intrinsic to the dynamics of the 20 industries themselves, at least when represented by panels almost coinciding with the whole population of persisting agents during the decade.

Null hypothesis of *1st-order stationary* Markov chain VS the alternative of *1st-order NOT-stationary* one (*Query a*)								
Matlab notation: $n\ e{-}m \equiv n{\cdot}10^{-m}$		Food (C10)	Textiles (C13)	Chemicals (C20)	Rubbers & Plastics (C22)	Machinery & Equipment (C28)	Inno SMEs	Inno Start-ups
Total Sales	Number of resident firms	4486	2728	2045	3286	8243	115	12
	Total number of proper degrees of freedom Δ_{df}	656	608	584	624	720	352	26
	Observed χ^2 total value	1405,717	2368,557	1553,495	3988,199	7595,798	343,588	37,971
	Observed MLRC final value	1327,430	2195,739	1354,963	3465,211	6594,503	320,983	29,453
	Null-hypothesis total Probability for χ^2 test	1,1214e-56	1,359e-205	4,9344e-89	0	0	0,6158	0,0609
	Null-hypothesis total Probability for MLR Criter.	8,5983e-48	5,039e-178	3,6578e-63	0	0	0,8810	0,2909
Total Assets	Number of resident firms	4497	2729	2048	3291	8251	115	14
	Total number of proper degrees of freedom Δ_{df}	568	512	456	520	632	272	30
	Observed χ^2 total value	1372,175	1132,880	756,149	1214,241	2611,963	286,269	23,164
	Observed MLRC final value	1256,699	1053,244	693,4596	1110,857	2462,692	268,672	26,129
	Null-hypothesis total Probability for χ^2 test	2,4336e-68	6,1082e-49	3,1501e-17	1,8682e-57	4,425e-238	0,2644	0,8085
	Null-hypothesis total Probability for MLR Criter.	4,8503e-54	1,0817e-39	4,3666e-12	5,5255e-45	1,042e-213	0,5456	0,6685
Total Debts	Number of resident firms	4497	2729	2047	3291	8251	115	14
	Total number of proper degrees of freedom Δ_{df}	608	616	552	600	696	304	27
	Observed χ^2 total value	1134,532	1075,021	713,433	1292,666	2842,067	286,658	36,637
	Observed MLRC final value	1076,517	1024,014	699,356	1231,243	2692,078	291,951	35,490
	Null-hypothesis total Probability for χ^2 test	2,7600e-34	1,9513e-27	3,8767e-06	7,7191e-53	2,914e-256	0,7549	0,1020
	Null-hypothesis total Probability for MLR Criter.	1,4415e-28	8,2118e-23	1,9623e-05	8,3674e-46	7,439e-232	0,6801	0,1269
Total Production Monetary Value	Number of resident firms	4493	2747	2050	3290	8275	115	12
	Total number of proper degrees of freedom Δ_{df}	664	608	592	600	704	304	27
	Observed χ^2 total value	1367,289	2218,114	1537,149	4076,975	8434,605	344,663	25,136
	Observed MLRC final value	1275,702	1912,866	1328,729	3487,154	7086,604	339,398	30,252
	Null-hypothesis total Probability for χ^2 test	5,5142e-51	1,500e-181	3,8432e-85	0	0	0,0540	0,5668
	Null-hypothesis total Probability for MLR Criter.	4,9259e-41	1,009e-134	1,6569e-58	0	0	0,0792	0,3029

Table 7.1: Here the Matlab notation has been adopted for 10-based powers, that is $n\ e{-}m \equiv n{\cdot}10^{-m}$, which is different from the scientific notation adopting operator E

Null hypothesis of *1st-order stationary* Markov chain VS the alternative of *2nd-order stationary* one (*Query b*)								
Matlab notation: $n\ e{-}m \equiv n{\cdot}10^{-m}$		Food (C10)	Textiles (C13)	Chemicals (C20)	Rubbers & Plastics (C22)	Machinery & Equipment (C28)	Inno SMEs	Inno Start-ups
Total Sales	Number of resident firms	4486	2728	2045	3286	8243	115	12
	Total number of proper degrees of freedom Δ_{df}	681	554	488	578	810	197	12
	Observed χ^2 total value	12465,689	6595,858	4756,903	9205,735	19394,581	718,262	15,509
	Observed MLRC final value	3109,865	2116,288	1381,859	2458,823	7920,317	302,999	14,029
	Null-hypothesis total Probability for χ^2 test	0	0	0	0	0	2,1248e-60	0,2148
	Null-hypothesis total Probability for MLR Criter.	8,965e-306	8,204e-181	2,2000e-86	1,462e-229	0	1,7841e-06	0,2988
Total Assets	Number of resident firms	4497	2729	2048	3291	8251	115	14
	Total number of proper degrees of freedom Δ_{df}	513	410	323	428	586	119	13
	Observed χ^2 total value	10594,030	2333,580	2478,411	6698,762	16410,739	164,212	7,366
	Observed MLRC final value	1739,013	1182,527	889,934	1509,952	4858,089	140,227	9,334
	Null-hypothesis total Probability for χ^2 test	0	7,886e-266	0	0	0	0,0038	0,8823
	Null-hypothesis total Probability for MLR Criter.	6,082e-133	5,2662e-76	1,6860e-54	1,811e-120	0	0,0894	0,7473
Total Debts	Number of resident firms	4497	2729	2047	3291	8251	115	14
	Total number of proper degrees of freedom Δ_{df}	579	565	471	543	730	158	8
	Observed χ^2 total value	11115,445	4669,356	3078,113	10554,980	15211,514	275,026	8,222
	Observed MLRC final value	2335,798	1609,985	1320,392	1917,433	6452,085	215,729	9,1358
	Null-hypothesis total Probability for χ^2 test	0	0	0	0	0	2,3472e-08	0,4121
	Null-hypothesis total Probability for MLR Criter.	5,889e-209	4,622e-101	1,3901e-81	1,931e-152	0	0,0016	0,3310
Total Production Monetary Value	Number of resident firms	4493	2747	2050	3290	8275	115	12
	Total number of proper degrees of freedom Δ_{df}	690	569	494	543	774	148	12
	Observed χ^2 total value	13260,085	6440,705	5997,978	7837,908	16029,691	331,177	11,349
	Observed MLRC final value	3243,698	2022,115	1479,501	2378,033	6654,608	206,718	13,964
	Null-hypothesis total Probability for χ^2 test	0	0	0	0	0	4,7169e-16	0,4992
	Null-hypothesis total Probability for MLR Criter.	0	1,253e-161	5,9187e-99	3,366e-227	0	0,0010	0,3030

Table 7.2: Here the Matlab notation has been adopted for 10-based powers, that is $n\ e{-}m \equiv n{\cdot}10^{-m}$, which is different from the scientific notation adopting operator E

We synthetically underline that the total number of the proper degrees of freedom has been every time reckoned via the restrictions on summations in Bickenbach and Bode (2003), without conditioning respect to their parameter m. The precise mathematical formulations are at page 369 (equations 6a and 6b, involving the sets A_i and B_i) and at page 371 (equations 8a and 8b, involving the sets C_i and D_i).

It is evident that the null hypothesis is unequivocally defeated by both the alternatives in the case of all variables for each ATECO sector, according to both the tests. It is proved by the fact that all values of the chi-squared statistics' ratio (or Pearson's ratio) and the MLRC's logarithmic statistics never contradict each other and are always greater, often considerably greater, than the corrected (or effective, or proper) overall number of freedom's degrees Δ_{df} for every one of the 20 possible matches between industries and variables excluding Innovative Start-Ups and Innovative SME (namely, for every ATECO data spreadsheet) at our disposal. It implies that occurrence probabilities of the null hypothesis in the two confrontations are almost infinitesimal. Let us remember that, when the empirical statistics of any chi-squared-type analysis computed over the data (or observed statistics' ratio) is greater than Δ_{df}, the null's occurrence probability is represented by the probability that the asymptotic χ^2 random variable may assume a value higher than the empirical statistics' ratio.

As regards the *Query a*, among such insignificant occurrence probabilities of the null, the largest ones have a 10^{-6} magnitude (χ^2) and a 10^{-5} magnitude (MLRC) and refer to the Total Debts of Chemicals' manufacture (or ATECO C20). For Rubbers & Plastics Manufactures (ATECO C22), as well as for Machinery & Equipment ones (C28), the null's occurrence probabilities computed for Total Sales and Total Production are zero. Regarding *Query b*, the magnitude of the largest probability, that the null hypothesis happens, is 10^{-54} for the MLRC applied to Chemicals' Total Assets and it is paired with a zero probability for the χ^2 test. Moreover, for the 20 matches except Textiles sector's (C13) Total Assets, probabilities in favour of the null equal to zero and in ATECO C28 the resulting probabilities are zero from both the tests.

So, it is legitimate to conclude that the 1st-order stationary Markov chain is a completely inadequate model (*as all the almost infinitesimal probabilities vouch for*)*, at least when the states are delimited by deciles, to describe the dynamics of four balance-sheet proxies, fundamental for any company, as Total Sales, Total Assets, Total Debts and Total Production Monetary Value, relative to the resident agents, during the decade 2006-2015, in five traditional and very important sectors of the Italian Economy, as Foods, Textiles, Chemicals, Rubbers & Plastics, Machinery & Equipment. An analogous verdict is found in Bickenbach-Bode (2003) about household incomes from a*

U.S.A. time series, by means of the same assessment criteria. Within the scope of a first-order Markov approach, χ^2 test and MLR criterion indicate not-stationarity as the additional property by far preferable (*inter alia, such a chain is the most advantageous one to visualise and study the date-by-date evolution of mobility indices based on transition matrices*)*; within the other scope of a time-homogeneous Markov approach, the same kind of analyses indicate the second one as the by far preferable chain order, similarly to Shorrock (1976). The possibility of systematic errors, originated by the Matlab device, is ruled out by the counterproof of Innovative Small & Medium Enterprises, because they generate appreciably different results albeit their core of incumbents is a good representation of a very-well-connotated, low-sized sub-population. Credibility of these investigations is guaranteed by size and thoroughness of every one of the 20 samples, consisting of many thousands of statistical units* (*much more numerous than the overall number of freedom's degrees*) *and almost overlapping with the real population of permanent businesses*. The present work strengthens, for five divisions of the Italian Market and four distinct financial items on a decade, the insights of previously cited authors on Anglo-Saxon incomes, as well as Hallberg (1969)'s ones for annual total volume of sales in gallons for milk products in Pennsylvania. Indeed, the 1st-order (I°) stationary (II°) Markov theory is grounded on too restrictive, necessary assumptions: property I°, meaning independence of one-step transition rates of any individual from its history, and property II°, stating independence from the time of one-step transition rates, are accompanied by the third tacit assumption of population homogeneity (i.e. independence of transition rates from agent's identity). Any national panel of thousands residing companies, close to the actual fixed population, is hardly able to fulfil them simultaneously. Then our inspections on the data suggest the convenience of attenuating those assumptions to raise up to an upper level of complexity, either through the introduction of pointwise dependence on time into the inference of transition probabilities for a first-order process or, wanting to maintain a stationary process, through the introduction of the simpler form of dependence on historical path for a chain, that is the set of second-order transition probabilities p_{khj}, where the jump from the state h at the time $t-1$ to j at time t is conditioned by the state k occupied at $t-2$.

Really, it is to be pointed out that this failure could not be entirely foreseeable *a priori* for financial quantities, akin to Total Assets and Total Debts, which are summations of many, more detailed and partial financial terms. Therefore, it would have been plausible, before any statistical assessment, the stationary 1st-order Markov paradigm to be sufficiently satisfactory, in force of possible compensation phenomena involving all together these partial sub-terms. But chi-squared-type tests demonstrate it does not happen so.

Textbooks about random variables and stochastic processes with applications, including the topic of Markov Chains, are countless: besides the classic Feller (1950), two more recent references are Grimmett-Stirzaker (2001) and Papoulis-Pillai (2002).

Comparing the Pearson's ratio to the MLRC logarithmic statistics for each of the 20 panels, we discover that the former is always higher than the latter for both the examined couples of the null hypothesis with its alternatives. Deviations are quite limited only for Total Debts in *Query a*, whatever the industry; in all cases of *Query b*, Pearson ratio's values are even multiple respect to the associated MLRC statistics' values (five times for C22 and C10 Total Debts, six times for C28). It also easy to note that, for most of the 20 matches, the parameter Δ_{df} is minor in the *Query b*'s "antinomy" than in *Query a*, excepting Total Sales, Total Debts and Total Production in the Machinery & Equipment manufacturing, as well as Total Sales and Total Production in the Food sector.

Now we need to comment the distinguished trends exhibited by the two groups of innovative companies (see the last two columns in Table 7.1 and Table 7.2) which are accounted for by the registers of InfoCamere and AIDA databases, even if they are not explicitly classified into the ATECO taxonomy. We warn the reader in advance that the number of extracted permanent start-ups is very scarce (12-14 units), thus our outcomes from applications of the χ^2 test and the MLR criterion on such an exiguous sample should be pondered with an extreme caution. In our analysis they are exclusively an example of misleading inference.

The problem of the size of Innovative SMEs' samples stays halfway between the situation presented by the five ATECO sectors and the Start-Ups one: the 115 extracted permanent SME firms are very less numerous than the thousands and thousands of incumbents in the "ATECO Five Ones", but it is however a

significant subset, around 38%, of the entire list of 302 registered SMEs in AIDA on one year at least, between 2006 and 2015. Incidentally, the population of Innovative Small & Medium Companies, by its very nature, is much less crowded nationwide than any traditional manufacture. Therefore, we believe that outputs from χ^2 and MLRC tests are substantially credible. Whatever the variable, the corrected number of freedom's degrees for the SME group exceeds the panel size, conversely to what happens in our traditional "ATECO Five", and it is approximately even twice, in case of *Query a*, the value of Δ_{df} in *Query b*. In both the inspected "antinomies" it is confirmed the χ^2 empirical statistics to exceed the MLRC analogous, i.e. the occurrence probability of the null hypothesis from the chi-squared test is smaller than the occurrence probability from MLRC: only for Total Debts in *Query a*, the chi-squared statistics is just a little lower than the MLRC one.

Relative to the antinomy of a 1st-order and time-homogeneous Markov chain against a 1st-order not-stationary one, Total Debts and Total Sales of Innovative SMEs disclose probabilities of more than 60%, consistently supporting the null for both the criteria, while Total Assets presents a minority 26,44% χ^2-test probability and a majority 54,56% MLRC probability on the side of the same null. Here it is found the unique, quite venial, incongruity between the tests in *Query a* and it could be interpreted by admitting the dependence of a component of SMEs' Total Assets respect to time (even if, at the current stage of the analysis, we are not able to identify such a component). Solely for the Total Production the 1st-order stationary chain is prominently disadvantaged respect to the alternative, presenting occurrences of 5,4% from the χ^2-test and 7,92% from the MLRC, which are, however, not inappreciable at all, unlike what happens for the 20 ATECO cases and are above the acceptance threshold of the 5% often used in statistical decisions. This preference highlights a meaningful pointwise time-dependence of the transition dynamics for SMEs' Total Production that is unknown to the other three SMEs' balance-sheet quantities, though it is a less strong dependence on time respect to the five manufactures. Looking carefully, an oddness emerges within *Query a*: *p*-values of both χ^2-tests are prominently uneven (albeit not negligible at all) for Total Sales compared to Total Production. Usually, in traditional ATECO sectors Total Sales and Total Production worth almost equally (every year, almost all revenues have been earned by selling what has been produced), but it is not so

among Inno-SMEs: evidently, their entire business includes a variety of activities and they are able to sell many services and consultancies being not part of the balance-sheet's Total Production. The latter discrepancy may derive from the fact that the group exists across the ATECO divisions, as well as from the constraints on size (Small & Medium) combined with their vocation to innovation. Dealing with the other antinomy of a 1^{st}-order and time-homogeneous Markov chain against a 2^{nd}-order stationary one, similarly to what happens for the five traditional industries, the null loses the challenge owing to very slight chances from a 10^{-60} magnitude to a 0,38% percentage. The sole incongruity among all the 24 cases in *Query b* concerns Total Assets, where an 8,94% MLRC chance, which is above the 5% acceptance threshold but also minority anyhow, is next to a negligible χ^2-test chance of 0,38%: diverging probabilities may rarely emerge from the two criteria when the sample size is modest.

A curious finding involves Total Sales and Total Production Value: despite they are always strongly correlated for all 'ATECO Five Traditional Ones', these two variables behave in dissimilar ways when Inno-SMEs are analysed.

The second Appendix A2 is constituted by sixteen tabulations where detailed elaborations distinguishing starting states, about the two hypothesis tests for the two antinomies, are reported from the workspace and organised in the special cases of the ATECO C10 division (Foods & Beverages) and the group of Inno-SMEs (eight tabs every one). Whatever the balance-sheet item, in *Antinomy a* of the Food Industry the second and the tenth inter-decile intervals, above all the tenth, sometimes along with the first cohort, give the smallest contributions to the overall statistical ratio and the null appear to be less unfit for them than for the other cohorts. Instead, the largest contributions to the overall statistics (ergo, the largest contributions to the rejection of the null) are given by intervals from the fifth to the eighth, especially by a pair among them, depending on the variable. In Foods & Beverages' *Antinomy b*, single-cohort contributions to the whole statistical ratio depend more markedly on the variable. The trend is that the third inter-decile stratum often yields a noticeable component, while a minor component comes from the tenth stratum. But there are some important exceptions, anyhow: the first cohort also allocates a big contribution to the Pearson's statistics, which becomes overwhelming for

Total Assets and Total Debts; generally, a modest component comes from the same first class to the MLRC statistics.

Shifting attention to Small & Medium Businesses, the distribution of the statistical ratio's partial terms over the ten inter-decile starting cohorts is quite diversified. Apropos *Antinomy a*, mainly the third and the fourth intervals flanked by the fifth one, provide the heaviest part of the whole statistics (i.e. the heaviest contribution to the refusal of the null) in case of Total Sales, Total Assets and Total Debts while for Total Production the distribution of the big terms is more uniform. Instead, whatever the item, the first and tenth inter-decile ranges allocate again the smallest components. A marked dependence on the variable of the single-class components' distribution is discerned also in SMEs Antinomy b, where no peculiar regularities are ascertained, saved the sixth inter-decile range often to yield a big contribution and saved the first inter-decile often to yield modest terms.

Finally, in regard to 1st-order Markov models applied to the basic proxies of assets, debt, production and sales, the patterns from 2006 to 2015 of our "ATECO Five" and the patterns of the outsider group of Small & Medium Enterprises in Italy have a propensity for opposite approximations: for the former ones, the time-inhomogeneity approach is surely the most appropriate in all the 20 considered combinations; for the latter, the stationary interpretation is better (Total production apart). In our opinion, the reason consists in the fact that each internationally classified, traditional manufacture is specifically characterized from a technological, an industrial and a commercial point of view; without any doubt, every traditional market is heavily conditioned by its large and very large companies; moreover, they belong to very complex and wider networks of businesses, interacting and competing among each other. In the period 2009-2015 of the Italian great economic uncertainty, these factors could have become remarkable in determining an absolute role of the time impact on firms' dynamics, not captured by a stationary model. While there are not large and very large protagonists inside the market of Innovative SMEs but, among them, all aspects of heterogeneity prevail, relative to the production system, output and customer typology, managerial and commercial strategies, since their principal common feature is precisely the attitude to innovation. So, pervasive heterogeneity in a SMEs 1st-order chain makes the time less important for the

statistical estimation of transition probabilities, in force of random compensation phenomena crossing the sample in any direction. It looks as if widespread heterogeneity in an evolving smaller sample encourages stationarity of the descriptive model.

Nevertheless, on the parallel side of time-homogeneous chains, for our four balance-sheet proxies in the epoch 2006-2015, both Innovative Small & Medium Enterprises and the five Italian industries of Foods & Beverages, Textiles, Chemicals, Rubbers & Plastics as well as Machineries have a generalized undisputed propensity in favour of the 2nd-order transitions against the 1st-order ones. The sense of such a conclusion is clearly that, for any kind of firm, the interpretation of its walk respect to a fundamental economic variable through a Markov paradigm, when time-inhomogeneity is overlooked, must include the influence of the preceding history, embodied at least by the last belonging stratum in the past, on the current jump.

A little more extensive discussion on the hypothesis tests' outcomes is reported in the following two paragraphs.

7.4 χ^2-type tests for Markov features: vulnerable to what degree?

Regarding hypothesis tests resting on the chi-squared distribution, some economists, who apply statistics to models' checking, suspect the χ^2-type tests have some weak spots.

Firstly, it is spreading the generic prejudice that they are prone to reject the null in the case of very large samples, whereas it is very likely that the test confirms the same null for small samples. For us it is a too simplistic conviction and some other statisticians assert when a χ^2-type dependability test refuses the null relatively to a dataset, then such a null is truly unsuitable for the current set: therefore, any model, anchored in that unacceptable null, cannot be a good model for the same dataset. We would like to add some other reflections. What matters is the relative size of the drawn sample compared to the whole population to be understood, not its absolute size in case of rejection: more precisely, what matters is how much representative the processed panel is compared to its source population, when a null hypothesis is overruled for the panel by the Pearson's test or by the MLRC. Moreover, χ^2-type tests

propensity to refuse the null may have not a logical sense if the compared statements are mathematically well-determined, as the Markov Chains' features involving distinct variants of the transition probability. Indeed, the chi-squared statistics of each of the two antinomies, *Query a* and *Query b*, we have surveyed is the ratio between the maximised likelihood functions of the current null and of its alternative (Anderson and Goodman (1957), at pages 97, 98, 100). Now, we could imagine swapping the roles of the null and of its alternative. The right statistical ratio for the new antinomy would be the inverse of the ratio between the maximised likelihood functions for the old antinomy: the same statement, previously ruled out in the old role of null hypothesis, after swapping would be inevitably excluded in the new role of alternative. Finally, there would be no change in the meaning of the assessment.

It seems more valid the second criticism hinged on sensibility of χ^2-type tests to little variations within very large samples (see Bergh (2015), Bertoni, Aletti et Alii (2918)). Bertoni, Aletti et Alii (2018) have blamed on the very big size of their incumbents' sample, consisting of hectares/parcels (in practice 639000 parcels for more than 743000 hectares), the failure, according to the Maximum Likelihood Ratio Criterion, of the 1st-order stationary chain as an adequate scheme for the sample evolution. For the moment, we do not know any scientific work connecting the χ^2 little-deviation-sensitivity to specific size-threshold of biggest samples. Nevertheless, it is reasonable to note that our largest panels, concerning the Machinery and Equipment C28-ATECO sector (8275 enduring factories in a decade for overall 18107 factories), are to a greatest extent smaller than the permanent parcels' panel of Bertoni and Aletti. Thence, we confide that the little-deviation-sensitivity has had a negligible impact on all χ^2-tests' *p*-values almost equal to zero in disfavour of Markov first-order stationarity, both for the five traditional manufactures in the chapter and *a fortiori* for the InnoSMEs group (115 residents per 302 firms in total).

7.5 Markov 1st-order stationarity and correlations in representative datasets of an industry

One of the needed requirements backstopping the statistical inference of a time-homogeneous chain, both at the first-order and at the 2nd-order, is

stochastic independence of individuals, constituting the fixed population, among one another, as it can be realised by looking at transformations linking to each other equations of the set 3.3 at page 91 in Anderson and Goodman (1957). But the stronger the correlations between individuals are, the more the requirement of independence is contradicted. Consequently, diffused and significant correlations among real individuals are a source of failure for the 1^{st}-order time-homogeneity when it is tested, onto their original real dataset, against 1^{st}-order time-inhomogeneity and 2^{nd}-order stationarity. In other words, diffused and significant correlations inside the dataset determine the fact that a 1^{st}-order stationary chain is not an appropriate model for these actual data, and alternative hypotheses in *Query a* and *Query b* must be preferred. As for our longitudinal panels of the five traditional ATECO manufactures, inner correlations are very likely for the most meaningful economic variables, so the latter argument is fairer than accusing some χ^2-statistics's flaws (which refer to different kind of data) of the failure. The work of Bertoni, Aletti et Alii (2018), once again, tells about strong correlations among parcels belonging to the same farm or to the same group of near (or similarly-behaving) farms. In like manner, *mutatis mutandis*, notable correlations may be realised with great probability inside every wide industry of an economically advanced country, and they can be divided into two kinds:

♦ horizontal or crosswise correlations, due to competition on the market between businesses of comparable size, at least;

♦ vertical or top-bottom correlations, due to subordination relationships, as the one from a prime contractor to a subcontractor, between businesses which are diverse in size.

It seems the 1^{st}-order time-inhomogeneity is more able to capture such correlations via explicit pointwise dependence on time connotating the conditional probabilities in the transition matrix. Moving to the 2^{nd}-order time-homogeneous chain, it wins against its 1^{st}-order analogue, although stationarity is a common trait for both, in case of all four variables and of the five ATECO division. Now, we must ask ourselves why it happens, despite correlations between actual enterprises exist in the market whatever the model. In this second case, we think correlations are captured by the dependence on the last two phases, before the present, of the history of the system, implemented via the second order of the chain, in lieu of being embodied by an explicit

pointwise temporal dependence inside first-order transition probabilities. In general, the order of any Markov Chain corresponds, by definition, to the number of backwards steps, before the present on the system's historical path, you want to include into conditional probabilities composing the model's transition matrix. The higher the Markov order is, the wider the fixed backwards width, respect to the current instant, accounted for in the past history at each step of the chain.

At an upper level, we also conceived (see the last paragraph in Chapter 6) single national manufactures akin to ecosystems, inhabited by economic subjects, weakly connected among one another. It is a compatible situation with the industry-specific correlations. However, if upper correlations between the national ATECO sectors exist, they are doing nothing but reinforce the lower industry-specific ones.

On the contrary, for the Inno-SMEs group from the AIDA archive (recall it is an outsider respect to the ATECO taxonomy), 1st-order stationary chain appears to be more acceptable, even when does not prevail in the case of Total Production and Total Assets, while competing versus the 1st-order not-stationary one, in its turn defeated for Total Debts and Total Sales. About this countertrend, we suspect there are some discriminating factors which weaken correlations within the group. The discriminants could be just those characteristics themselves, defining Innovative SMEs as a little and strictly-specified sub-population of firms. More precisely:

♦ because of the tiny size of their subpopulation, Innovative SMEs are dispersed in the Italian economy; in a way, we could say such subpopulation is "diluted" inside the Italian industrial scenario to an extent that horizontal correlations due to competition come out heavily lessened;

♦ their keen and distinctive vocation for innovation might facilitate a sort of isolation of the subpopulation, hindering vertical correlations with enterprises belonging to more classic ATECO divisions and promoting direct interactions with their own market niche (it must be reminded that innovation rate is quite scarce for the most of Italian firms).

After conjecturing that a stationary chain is not a good model for changes in any industry whose firms actually are strongly correlated, and that inside Inno-SMEs group that correlations between statistical units are absent (or they are very weak), it has just to be expounded why, also in the latter subpopulation,

both the Pearson's test and the Maximum Likelihood Ratio Criterium assign in practice an 100% *p*-value to Markov 2nd-order stationarity in contrast with the 1st-order stationary chain (*Query b*, the sole exception is the MLRC *p*-value of 8,94% already cited for Total Assets), as it happens for all traditional "ATECO Five". Namely, why stationarity seems not to be, only within the second antinomy, a disadvantage of the chain respect to the subpopulation's data. The reason may be the second order itself. Regardless of existence and force of stochastic correlations among agents in any industry, hence regardless the model stationarity, previous history (which is implemented into the Markov chain via an order higher than the first) is a helpful determinant to better understand and profile industrial structural evolution.

Ultimately, there is no doubt that all argumentations proposed in the last two paragraphs should be matter of further, meticulous quantitative investigations.

Chapter 8
Data analysis, part 3: economic mobility in some Italian industrial sectors

8.1 A little preamble on mobility measures

Despite the question of the reliability of 1st-order stationary Markov model to forecast observed frequencies distribution of the fundamental quantities in economics (gradually enlightened in Chapter 1 and Chapter 2; discussed in the background of Italian data in Chapter 7), there are no reason to deny that a *rigorous statistical estimation of the first-order not-stationary transition matrix is, however, a compelling collection of information dealing with the evolution of agents' placement among states, from any date to the next one. Then, all such information* on changes from a state to another, contained in the transition matrix, *might be summed up into a functional of the matrix itself, constituting an interpretation of the mobility inside the system.* Incidentally, 'transition' and 'mobility' are synonyms in the common language. Transition matrices are the theoretical analogue, since every row must sum to 1, of mobility tables that were object of a wide reflection in social sciences, after the 1955 work of Prais, by Bartholomew (editions from 1967 to 1982), Boudon (1973), Bibby (1975). Though, they waived to discuss which absolute properties should connotate any measure of agents' movement among classes (i.e. the mobility): such a task was faced organically in Shorrocks (1978), where key points of the axiomatic theory of any mobility index is fixed for a discrete temporal parameter and a discrete space of states.

Really, in the socioeconomic theory it is possible to think up *many different notions of mobility* (applicable to households among levels of incomes or wealth, to workers among occupations, to firms respect to total assets or respect to the number of employees etc.). They refer to more than one axiomatic theory, because there exists a variety of independent notions of 'perfect mobile society/economy', and each of them inspires a lot of diverse mathematical measures. Remember that, right in the first paragraph of this

chapter, we have already cited the discussion on social mobility by Prais, without using a dedicated index, and the mobility index by Adelman, not directly established on the Markov transition matrix, but on mean permanence durations of firms inside the states and on independence of transitions from the starting state. Fields (2008), which give a good synthetic exposure about this topic, indicates 6 distinct concepts: time-independence, absolute movement, quantile movement, share movement, non-directional movement (or 'flux'), directional movement and, finally, longer-term equalizer. This author says at least 20 mobility measures derives from them in the scientific literature and from 2008 up to now they have surely become more numerous. *Mobility indices* are the suitable tools for such measures. Any mobility index is a proper function associating only one scalar (which should be normalized inside the real interval [0,1]) to a variation, occurring between two distinct point in time, in the distribution of the values expressed for a variable by a population or a sample of individuals. Its task is to summarize movements of the agents, interpreted as whole phenomenon, during a fixed time interval. Fields and Ok (1999) is an extensive survey (although inevitably not exhaustive) on income mobility, from which most of the general ideas can be rightfully adopted for other persons' or firms' economic variables. The same goes for the two slightly shorter articles Fields and Ok (1996) and Fields (2000). They emphasize mobility indices defined as combinations of modifications with time of the rank position for every agent inside the population (or combinations of functionals on such modifications), but it is not the sole option.

Another type of mobility measure thoroughly investigated is the class of indices defined precisely over the space of transition matrices synthesizing raw data and the 1978 paper of Shorrocks, mentioned above, is an early reference. While presenting the fundamental properties required by an index of mobility, he observed that they become inconsistent, if regarded as a set of axioms, due to a conflict between monotonicity (related to the intuition that the index should grow when the off-diagonal elements of the transition matrix become larger) and the definition adopted for a perfect mobile structure (that matches the value '1' for any index). So, he conjectured that this conflict originates from matrices very unlikely to arise in practice and proposed to remedy by ignoring them. In truth, such a conflict might also be avoided either by modifying the definition of perfect mobility; or by associating the value '1' of

a mobility measure to any set of changes, inside the sample (population), realizing the *total maximal movement*, without excluding *a priori* any matrix. Then he tried to tackle the issue of the comparison between different values of the same index, computed for various transition matrices referring to a unique variable on different periods of time (as it may happen when managing two or more databases for the same research). Shorrocks's solution was, in order to make the description of mobility free from the influence of time, the introduction of the stronger axiom of period invariance, after making explicit the dependence of the matrix on time. Finally, he disclosed two new period-invariant measures: one is a functional of the determinant of the matrix, the other rests on the asymptotic half-life in a 1st-order stationary Markov chain (asymptotic half-life in its turn is linked to the speed of convergence towards the equilibrium distribution).

Let us remind that all indices **in *MarkovInfer_MobInd.m* pertain to mobility exclusively in an intragenerational sense** (as it has been revealed in advance at the beginning of Chapter 1), which is the most manageable one when whoever wants to study the mobility of companies inside a specific industrial sector. In the case of businesses, investigations on mobility in an intergenerational sense would be more demanding and ambiguous since it is not possible to determine in a univocal manner, or predominantly at least, a relationship of the kind 'parents-offspring' for businesses to the same extent as for human people (workers, households, stakeholders etc.). Potential definitions of 'parent' and 'son' among businesses are more than one, owing to differences among mergers/break-ups and acquisitions; they might be forged by false bankruptcy and simultaneous re-denominations; they might give some troubles, for instance the systematic and immediate 'parents' death while its 'son' was just coming into the world (which occurs very rarely in the human world). Besides, it would be demanding to precisely select such different parental relationships within businesses' databases.

8.2 Some hints for the axiomatic approach to mobility indices

The axiomatic characterization of any mobility measure $I(P)$ resting on the transition matrix P is in Shorrocks (1978) and should ideally consist of the properties below.

Just remember that, by the definition of transition matrix, the element of P located at the i-th row and the j-th column is the probability that the state $s(t+1)$ of an individual of a system (a sample, a whole population) at the instant $t+1$ is j, provided that it is i the state $s(t)$ occupied by the same individual at the instant t :

$$[P]_{ij} = p_{ij} \equiv p[\ s(t+1) = j \mid s(t) = i\ ],$$

where $p[\cdot \mid \cdot]$ is the transition probability distribution (it is a conditional random distribution).

• {N} (Normalization) It is a continuous real function $I(\cdot)$ defined over the space of all $k \times k$ transition stochastic matrices $\mathbb{P}_{k \times k}$ towards the real segment of normalization [0,1] ($\Re$ is a symbol for the real line):

$$I(\cdot) : \mathbb{P}_{k \times k} \to \Re \ , \ \forall P \in \mathbb{P}_{k \times k} \quad \mathrm{P} \to I(P) \in [0,1].$$

P is a stochastic matrix in the sense that $\Sigma_{j=1,2,..,k}\, p_{ij} = 1 \ \ \forall i = 1,2,..,k$.

• {M} (Monotonicity) It is monotone: $\quad P > P' \Rightarrow I(P) > I(P')$

where $P > P'$ means that $p_{ij} \geq p'_{ij} \ \ \forall i \neq j$ and $p_{ab} > p'_{ab}$ only for some couples (a,b) such that $a \neq b$.

• {I} (Immobility) Its value for the identity matrix must correspond to its minimum 0 meaning immobility. From monotonicity derives the association of the identity matrix $Id_{k \times k}$ of $\mathbb{P}_{k \times k}$, representing the complete absence of jumps between the states bolstering any transition matrix, to the minimum index value: $I(Id) = 0$.

• {PM} (Perfect Mobility) It is worth its maximum 1 in correspondence of any perfectly mobile matrix. If $\boldsymbol{u}$ and $\boldsymbol{x}$ are two k-component row vector, $\boldsymbol{u} = (1,1,..,1)$ and all components of $\boldsymbol{x}$ sum to 1 (in other words $\boldsymbol{x} \cdot \boldsymbol{u}' = 1$) then $P = \boldsymbol{u}' \cdot \boldsymbol{x}$ is a perfectly mobile transition matrix and:

$$P = \boldsymbol{u}' \cdot \boldsymbol{x} \ \Rightarrow \ I(P) = 1.$$

A "perfectly mobile system", recapped via any above-cited product $\boldsymbol{u}' \cdot \boldsymbol{x}$, is conceptually defined as the one where the conditional probability of reaching whatever final state depends solely on the latter but does not depend on the starting state. A perfectly mobile matrix $P = \boldsymbol{u}' \cdot \boldsymbol{x}$ consist of k rows identical to the vector $\boldsymbol{x}$, thus $\forall j = 1,2,..,k$ the elements of its j-th column are all equal to the vectoral component x_j , albeit they are different from the elements of any other

column. It must be underlined that the vector $\boldsymbol{x}$ is not unique, it could be identified with the equilibrium distribution of the underlying 1st-order Markov chain, even if it is not necessary at all.

Both last properties admit a stringent version to set aside extreme index values to the identity matrix and to the matrix of perfect mobility:

- {SI} (Strong immobility) The condition holds when

$$I(P) = 0 \Leftrightarrow P = Id$$

- {SPM} (Strong perfect mobility) The condition holds when

$$I(P) = 0 \Leftrightarrow P = \boldsymbol{u}' \cdot \boldsymbol{x}$$

where $\boldsymbol{u}$ and $\boldsymbol{x}$ are the vectors introduced in {PM}.

Further stronger properties can be uncovered for specific measures (Maximum Mobility, Strong Maximum Mobility, Independence from irrelevant classes, Equivalence when rows are symmetric respect the main diagonal in Alcalde-Unzu, Ezcurra, Pascual (2006) are an example).

Unfortunately, the axiomatic theory of mobility is not as generous as the axiomatic theory of probability. *De facto*, there is no index capable of satisfying all requirements of the previous axiomatic set. For instance, at the moment we do not know if the Alcalde-Unzu Family of Normalised Indices are maximised to 1 in case of perfectly mobile transition matrices, despite the extra attributes mentioned just above. Only {I} has a universal validity whilst there may be incompatibilities between {N}, {M}, {PM} for some particular matrices. In Shorrocks (1978) it is stated the basic conflict should be between {PM} and {M}; a discussion about this issue is developed at page 1016. As regards the notion of mobility we are thinking about, a proper mathematical gauge must be chosen also according to the axiomatic features we are willing to abandon. Moreover, the more and more sophisticated mathematical definitions of some recent indices have induced to make more complex the requisites of the mobility axiomatic theory. The Directional Index by Ferretti and Ganugi, in addition to interpreting mobility as a combination of the movements of agents, takes also into account orientation of movements; it is normalised into the real segment [−1, +1] instead of into the real [0,1], therefore it obey to a peculiar axiom of weak immobility and the axiom of perfect mobility has to be split into two 'sides', relative to jumps on the

positive direction (i.e. towards higher strata) and to movement on the negative one (i.e. towards lower strata).

From empirical observations of academics' repeated researches, two principles arise concerning mobility tendency to increase, easy to be explained:

– *principle of inflation according to the temporal spans' dilation*;

– *principle of inflation according to the space splitting*.

We conjecture the mobility increase in the former principle is unbounded whereas mobility increase in the latter principle is asymptotic respect to the maximum value admitted for the index due to normalization.

The former is reported in Shorrocks (1978) and, for incomes, in Atkinson with Bourguignon and Morrisson (1992). After fixing a starting time, if it is considered an epoch longer and longer, displacements of any statistical units among states will be more and more likely. At the end of the epoch, it will be more and more likely to watch more and more translocations into a state different from the one of departure. Consequently, the number of units drifted from the initial state to another final one during the epoch will become greater and greater, whatever the couple of the starting state with the state of arrival. Values of transition probabilities of the kind $p_{ij,i\neq j}\ \forall i \neq j$ will become higher and higher at the expense of the probabilities of staying $p_{ii}\ \forall i$, hence estimated mobility should increase.

The latter is fully justified when the model, adopted to portray an evolving system, is sensitive exclusively to shifts between disjoint finite states partitioning the space. Then, the prospect of a prefixed-length movement to be detected within the space depends on the extent of the departure state. The more numerous are disjoint classes splitting the same space, the minor is the average extent of such classes, the higher is the probability that the model is able to view a certain prefixed-length movement of any agent in the space as it was a shift from the starting state to a diverse one.

The mobility inflation with the temporal span dilation is linked to the problem of comparing (with the purpose of making rankings) some values of the same index grounded on distinct transition matrices which in their turn refer to different time intervals, because the surveyed historical age varies with the data. The problem has been considered irrelevant in studies about intergenerational mobility (changes of the educational level, of the belonging social class from fathers to sons), for which the periodization is grounded on

the idea of generation, whereas it is fundamental in comparisons between values of an intragenerational index (changes of income strata the same families belong to). In practice, we avoided any troubles in this regard referring to the same electronic archive, AIDA, where observations are annually detected during a decade, for any industrial sector and any variable throughout the Italian country. In theory, the problem could be overtaken when the index satisfies the further properties of period consistency and period invariance reported below.

• {PC} (Period Consistency) If P and Q are transition matrices of 1st-order Markov chains describing two different systems, the condition consists in:

$$I(P) \geq I(Q) \Rightarrow I(P^n) \geq I(Q^n) \quad \forall n = 1, 2, 3, \ldots.$$

• {PI} (Period Invariance) After explicitly introducing, into the mathematical definition of the index, in a suitable manner the time period T to which the transition matrix refers, the shape of the mobility proxy becomes $I(P\,;\mathrm{T})$. It is period invariant when:

$$I(P,\mathrm{T}) = I(P^n\,;\, n\mathrm{T}) \quad \forall n = 1, 2, 3, \ldots.$$

Although are {PC} and {PI} desirable from a purely theoretic point of view, they are also too restrictive: beside the risk of excluding the most of mobility measures already devised, the few permitted measures could not to work very well in an empirical sense, creating some difficulties in statistical analyses on actual data, e.g. mobility values becoming almost zero too fast. It is an example what happens for the Asymptotic Half-life Index (see the last paragraph of discussion in the current chapter and tabulations in Appendix A.3). An alternative way to produce mobility rankings, in case of systems whose data are yielded from epochs of different lengths, is to apply Continuous-Time Markov Chains to the system dynamics.

8.3 Which Transition-matrix-based mobility indices are in MarkovInfer_MobInd.m

In the previous chapter we have verified with scientific rigour by two distinct chi-squared type tests that the null hypothesis of 1st-order time-homogeneous Markov chain is rejected in favour of both the alternatives, the 1st-order not-stationary chain and the 2nd-order stationary one. In order to go ahead in the analysis of the outputs elaborated by *MarkovInfer_MbInd.m* towards the topic

of mobility measures, it is better to choose one between the winning models. As far as we know, any statistical test does not exist yet to directly evaluate respect to each other the reliability of those two alternatives. It would be very arduous and long-lasting to redefine the nine transition-matrix-based mobility indices relatively to the 2nd-order time-homogeneous chain as a subtended model for a *m*-classes Markov system, because it is necessary to transform the 2nd-order chain into a pseudo 1st-order one, as well as the 2nd-order transition probability p_{jkh} into a formal 1st-order conditional probability $p_{(j,k)\to(k,h)}$ referring to an *m*-classes system. In such a reformulation, one-step transitions between consecutive stages are no longer mono-period transitions between simple states. It would be needed to begin considering a new kind of jumps involving consecutive compound states, (j,k) and (k,h), every one of them composed by two simple classes, j and k respectively from the time $t-2$ until $t-1$, for the starting stage; then k and h respectively from the time $t-1$ until t, for the ending stage. So, after this reconfiguration of the chain, every feature of the axiomatic definition of a mobility index for Social and Economics Sciences should be revisited. Moreover, the restyled transition probabilities $p_{(j,k)\to(k,h)}$ would sort themselves out into a cumbersome transition matrix sized $m^2\times m^2$, where each element corresponding to an impossible jump, like $(j,k_1)\to(k_2,h)$ when $k_1 \neq k_2$, is worth zero. In a similar context, any scientific overview on mobility through the 2nd-order stationary chain would turn out to be much more intricate: therefore, we decided to study economic mobility through the two type of simple Markov chains.

The transition-matrix-based mobility measures, implemented in the simplest 1st-order Markov frameworks, the time-homogeneous chain as well as the not-stationary one, calculated through our Matlab device, are reported in the following list.

I1) the Normalized Directional Index (Ferretti and Ganugi (2013))

I2) the family of Normalized Relative Indices of Mobility as a Movement (Alcalde-Unzu, Ezcurra, Pascual (2006))

I3) the Trace Index (Shorrocks (1978))

I4) the Exponentiated, or Amplified or Generalized, Determinant Index (Shorrocks (1978))

I5) the Index of Predictability (Parker and Rougier (2001))

I6) the 2nd Bartholomew Index (Bartholomew (1982))

I7) the Second-Eigenvalue Index (Sommers and Conlisk (1979))

I8) the All-Eigenvalues Index, or Spectral Index (Sommers and Conlisk (1979))

I9) the period-invariant Asymptotic Half-life Index (Shorrocks (1978)).

We avoided considering any index involving not only transition probabilities but also the equilibrium distribution of a 1st-order stationary Markov chain, akin to some mobility proxies in Bartholomew (1982) owing to the proved inadequacy of the Markov first-order time-homogeneity to capture the genuine dynamics of classic Italian manufactures (see Chapter 7, paragraph 7.3).

A coincise but exaustive discourse about a variety of concepts of mobility, along with the corresponding mathematical definitions as functions of the transition matrix, can be found in Ferretti (2012). More extensive surveys and researches, dealing with many different paradigms for the measurement of mobility, with an exclusive emphasis on the income variable but whose contents are still true or can be translated/adapted (and could be examined) when moving towards diverse economic variables, have been produced by Fields and his coauthors. Some examples, proposing either a theoretical approach or an empirical one, are: Fields (2000), Fields (2008), Fields and Ok (1996), Fields and Ok (1999).

In the first part of the Appendix A3 all estimations of the nine measures, according to the two Markov approaches, relative to their mono-period version (i.e. considering transitions between consecutive dates, or one-step transitions, from $t-1$ until t, whatever the year $t-1$ inside the age 2006-2014) are organised in tabulations.

For every measure and for both the Markov models, the aggregated-in-time, or time-compound, variant has been also calculated starting from the first registered date inside the 28 datasheets: for instance, as regards the 24 combinations between an economic variable and an ATECO manufacture or the group of Innovative Small & Medium Enterprises, indices has been computed on all the time spans from the year 2006 until each of the subsequent dates 2006 + n (up to covering the whole decade 2006-2015 in the data spreadsheets). These variants are indicated by the progressive initials from AI1 up to AI9 in the same order of the list cited just above. Such time-aggregated variants of the nine indices are displayed in the tables of the second part of

Appendix A3. A careful explanation on how the simple stationary Markov chain does work, and about aggregation respect to the time involving the stationary mono-period transition matrix, can be read in the pioneering paper Prais (1955), albeit social mobility is treated here instead of mobility regarding economic proxies.

8.4 Mobility as an oriented and a not-oriented movement

The Directional Index, whose mathematical definition is (Ferretti and Ganugi (2013), at page 407)

$$I_{\omega,v}(P) = \sum_{i=1}^{k} \omega_i \cdot \sum_{j=1}^{k} p_{ij} \cdot sign(j-i) \cdot v(|j-i|) \quad [8.1]$$

recalls the concept of mobility as an overlapping of oriented movements of the statistical units within the domain of a fixed variable. It consists of several terms.

• The factor $sign(j-i)$ is the *signum function* (which is worth −1 if $j-i<0$, is zero if $j-i=0$ and is equal to +1 if $j-i>0$) evaluated on the jump's magnitude $(j-i)$ for individuals moving from i to j. It is the peculiar feature of $I_{\omega,v}(P)$ which is the sole mobility index gauging at the same time the degree of mobility and the prevailing orientation of movements. Among firms moving from i to j, some of them have moved via an upsizing if $j>i$, the others via a downsizing if $j<i$. The corresponding contribute to the overall mobility, due to the factor '*sign*', will be respectively positive and negative: if (and only if) the most part of jumps is an upsizing, firms tend on average to improve their condition, so it is $I_{\omega,v}(P)>0$; otherwise the index will assume negative values.

• From [8.1] we see the index is a weighted arithmetic mean of each value of the partial directional mobility $\sum_{j=1}^{k} p_{ij}\, sign(j-i)\, v(|j-i|)$ from every single state, due to individuals starting from the i-th interval ($i = 1,.., k$). Every firm, starting from the i-th state, has a potential mobility related to the conditional probabilities $\{p_{i1},.., p_{ik}\}_{i=1,\dots,k}$. The weights are $\{\omega_i\}_{i=1,\dots,k}$ such that $\omega_i \geq 0$ for every i and $\sum_{i=1}^{k} \omega_i = 1$. We stress the index in [8.1] permits many different choices of $\{\omega_i\}$, relevant in a variety of respective situations: someone, interested in the sole mobility due to shifts from the first state, could obtain the corresponding measure simply by fixing $\omega = [1, 0, ..., 0]$.

• The function $v(\cdot)$ is non-decreasing, such that $v(0) = 0$ and it serves to measure the magnitude of jumps from i to j, for every couple $i, j = 1,.., k$. Similarly to the coefficients $\{\omega_i\}$, $v(|j - i|)$ can be seen as a weight which assigns a larger role in the whole mobility to individuals making larger jumps. Indeed, suppose indeed that, for a given matrix P, it holds $p_{12} = p_{13}$: if $v \equiv 1$, firms moving from $i = 1$ to $j = 2$ and firms moving from $i = 1$ to $j = 3$ have the same impact on mobility. Instead, by setting $v(|j - i|) \equiv |j - i|$, we assume that jumps from a given state i to the state $i+2$ have a double weight in the mobility with respect to any jump from i to $i+1$. Otherwise, it often happens that the chosen size classes become exponentially larger and larger, and movements from $i = 1$ to $j = 3$ are very harder than movements from $i = 1$ to $j = 2$: one of the suitable choice may be $v(|j - i|) \equiv e^{|j-i|} - 1$. For sake of concision, the notation: $V_{ij} \equiv \tilde{v}(j - i) \equiv sign(j - i) \cdot v(|j - i|)$ could be used.

• If other observation dates about the process $X(t)$ had been registered between the transition extremes t_0 and t_1 to which the states i and j are referring, the operator $\tilde{v}(j - i)$ might also depend on the intermediate smaller drifts associated to the middle states. But it is not contemplated by the original definition of the Directional Index: it is a more sophisticated modulation, requiring verifying the new version of the Index as regards all axiomatic properties characterizing any mobility measure. It would be the aim of a further work.

Equation [8.1] can be reshaped in a more compact form by vectors and matrices. Indeed, if the weights $\{\omega_i\}_{i=1,\dots,k}$ are the ordered components of the row vector ω, if $V_{ij} \equiv sign(j - i) \cdot v(|j - i|)$ is the element positioned in the i-th row and the j-th column inside the matrix V, if $Y_i \equiv (P \cdot {}^{T}V)_{ii}$ is the i-th entry of the row vector Y composed by all the elements on the principal diagonal of the product $P \cdot {}^{T}V$, then the definition [8.1] becomes:

$$I_{\omega,v}(P) = \sum_{i=1}^{k} \omega_i \cdot (P \cdot {}^{T}V)_{ii} = \omega \cdot Y$$

It is computable by a shorter list of instructions in any environment capable to handle vectors and matrices, like MATLAB.

It is possible to demonstrate that $I_{\omega,v}(P)$ assumes its values in the closed and bounded interval $[m(\omega, \tilde{v}, j=1), m(\omega, \tilde{v}, j = k)]$ where $m(\omega, \tilde{v}, j=1) \leq 0$ and $m(\omega, \tilde{v}, j = k) \geq 0$ do not depend on P and:

$$m(\omega, \tilde{v}, j = 1) = \sum_{i=1}^{k} \omega_i \ \tilde{v}(1 - i) \qquad [8.2]$$

$$m(\omega, \tilde{v}, j = k) = \sum_{i=1}^{k} \omega_i \ \tilde{v}(k - i) \qquad [8.3]$$

(Ferretti and Ganugi (2013) p. 408). Such two boundaries directly depend on the oriented jumps' modulation $sign\,(j - i) \cdot v\,(|j - i|)$, on the weights $\{\omega_i\}$ of the starting states and, respectively, on the lowest and the uppermost destination classes of the underlying state space. Hence, they are conditioned by the researcher's settings about the underlying space as well as about the directional mobility. However, in the rest of the chapter they could be briefly denoted as $m_1 \equiv m(\omega, \tilde{v}, j = 1)$ and $m_2 \equiv m(\omega, \tilde{v}, j = k)$. The directional index is equal to m_2 (alternatively m_1) if and only if maximum upsize (alternatively downsize) happens, that is when every row of the transition matrix P is equal to [0, 0, …, 1] (alternatively [1, 0, ..., 0]) and every firm directly jumps in the best (resp. worst) size class. The normalized version of the index $I_{dir}\,(P)$ obtained dividing $I_{\omega,v}\,(P)$ by the absolute values of its maximum and minimum:

$$I_{dir}(P) = \begin{cases} \dfrac{-1}{m_1} I_{\omega,v}(P) & \text{if } I_{\omega,v}(P) < 0 \\ \dfrac{1}{m_2} I_{\omega,v}(P) & \text{if } I_{\omega,v}(P) \geq 0 \end{cases} \qquad [8.4]$$

Normalization is always feasible because m_1 and m_2 cannot be simultaneously null. By definition, I_{dir} is equal to $+1$ in the best case of maximum upsize and -1 in the case of maximum downsize. As a consequence, the normalized index can be thought as a percentage of the extreme cases: e.g., $I_{dir} = -0.5$ means that the mobility in the sample is towards downsizing and its intensity is a half than the case of maximum downsizing.

It might be of some use, to describe a little more accurately the directional mobility phenomenon, distinguishing in the sample the units moving to an upper class from the ones going back to a lower class and the others standing still.

Let us introduce three matrices whose rows do not sum to 1 (they are not stochastic, though their elements are random variables), differently from the transition matrix P (whose elements are $[P]_{ij} = \{p_{ij} \ \ i,j = 1,2,.., k\}$):

– the upper part of the transition, P^{UP}

$$[P^{UP}]_{ij} = p_{ij} \text{ if } i < j \text{ and } [P^{UP}]_{ij} = 0 \text{ if } i \geq j$$

– the lower part of the transition, P^{DN}

$$[P^{DN}]_{ij} = p_{ij} \text{ if } i > j \text{ and } [P^{DN}]_{ij} = 0 \text{ if } i \leq j$$

– the immobility part at the initial state i, $P_{(i)}$

$$[P_{(i)}]_{ij} = p_{ii} \text{ if } i = j \text{ and } [P_{(i)}]_{ij} = 0 \text{ if } i \neq j$$

The next decomposition simply holds:

$$P = P^{UP} + P^{DN} + P_{(i)}$$

so, by substituting it into the definition of the directional mobility index, equation [8.1] is decomposable as:

$$I_{\omega,v}(P) \equiv I_{\omega,v}(P^{DN}) + I_{\omega,v}(P_{(i)}) + I_{\omega,v}(P^{UP}) \equiv I^{DN}_{\omega,v}(P) + I^{UP}_{\omega,v}(P)$$

that is

$$I_{\omega,v}(P) \equiv \sum_{i=1}^{k} \omega_i \sum_{j<i,j=1}^{k} p_{ij} \cdot (-1) \cdot v(i-j) + \sum_{i=1}^{k} \omega_i \sum_{j>i,j=1}^{k} p_{ij} \cdot v(j-i)$$

where it is zero $I_{\omega,v}(P_{(i)}) \equiv \sum_{i=1}^{k} \omega_i \cdot p_{ii} \cdot sign(i-i) \cdot v(|i-i|)$ the contribution of all individuals lingering on the starting state after the transition of the others, due to $sign(i-i) = 0$. So

– the *upward component of the directional index*

$$I^{UP}_{\omega,v}(P) \equiv \sum_{i=1}^{k} \omega_i \sum_{j>i,j=1}^{k} p_{ij} \cdot v(j-i) \geq 0 \qquad [8.5]$$

is a quantity summarising the mobility of the part of the population jumping to any final state *upper* than the starting one;

– the *downward component of the directional index*

$$I^{DN}_{\omega,v}(P) \equiv \sum_{i=1}^{k} \omega_i \sum_{j<i,j=1}^{k} p_{ij} \cdot (-1) \cdot v(i-j) \leq 0 \qquad [8.6]$$

measures, inside the index, the mobility of the part jumping to any final state *lower* than the starting one.

When $\omega_i \equiv d_i$ and $v(|j-i|) \equiv 1$, the upward component and the absolute value of the downward component of the Directional Index coincide respectively with the Upward $I^{UP}_{BM}(P)$ and the Downward $I^{DN}_{BM}(P)$ Indices of Bourguignon and Morrison (Bourguignon and Morrison (2002)):

$$I_{BM}^{UP}(P) \equiv \sum_{i=1}^{k} d_i \sum_{j>i,\, j=1}^{k} p_{ij} \quad \text{and} \quad I_{BM}^{DN}(P) \equiv \sum_{i=1}^{k} d_i \sum_{j<i,\, j=1}^{k} p_{ij}$$

where d_i is the fraction of individuals moving from the state i.

Obviously, the normalization of the index affects the upward and downward components as:

$$I_{dir}^{UP}(P) = \begin{cases} -I_{\omega,v}^{UP}(P)/m_1 & \text{if } I_{\omega,v}(P) < 0 \\ I_{\omega,v}^{UP}(P)/m_2 & \text{if } I_{\omega,v}(P) \geq 0 \end{cases}$$

$$I_{dir}^{DN}(P) = \begin{cases} -I_{\omega,v}^{DN}(P)/m_1 & \text{if } I_{\omega,v}(P) < 0 \\ I_{\omega,v}^{DN}(P)/m_2 & \text{if } I_{\omega,v}(P) \geq 0 \end{cases}$$

(remember it is $m_1 \equiv m(\omega, \tilde{v}, j=1)$ and $m_2 \equiv m(\omega, \tilde{v}, j=k)$).

Actually, few years before the Directional Index was invented by Ferretti and Ganugi, Alcalde-Unzu along with his co-authors had already presented a family of indices defined by a nonlinear combination of "distances". Such distances are between the hypothetical situation of immobility, represented by the identity matrix (whose element in a whatever position is the Kronecker-delta symbol δ_{ij}), and any element positioned in the i-th row and in the j-th column inside the transition matrix (namely, the conditional probability of a jump from whatever initial state i to any final state j). They have named such a family "*relative indices of mobility as a* (*not oriented*) *movement*" and it is subtended by the notion of mobility as a combination of movements whose intensity is essential, whereas their orientation is overlooked. It is studied in detail in Alcalde-Unzu, Ezcurra, Pascual (2006).

A mobility index belongs to the family $I^{N}{}_{\omega,v,\alpha}(P)$ of "Normalized Relative Indices of Mobility as Movement" if there exists a vector $\omega \equiv (\omega_1, \ldots, \omega_k)$ where $\omega_i \in \Re$, $\omega_i \geq 0$ and $\Sigma_{j=1\ldots k}\ \omega_j = 1$, there exists v: $\Re \to \Re$ a scalar strictly-increasing function , $v > 0$, along with a real number $\alpha \geq 1$ such that, $\forall P \in \mathbb{P}_{k\times k}$

$$I^{N}{}_{\omega,v,\alpha}(P) = \frac{1}{Z} \sum_{i=1}^{k} \omega_i \left[\sum_{j=1}^{k} | p_{ij} - \delta_{ij} |^{\alpha} \cdot v\,(\,|i-j|\,) \right]^{1/\alpha} \qquad [8.7]$$

$$Z \equiv \sum_{i=1}^{k} \omega_i \left[v(0) + \max[v(i-1) \cdot v(k-i)] \right]^{1/\alpha} \qquad [8.8]$$

$1/Z$ is the scalar factor whereby the index is normalized into the real interval [0,1]. The parameters $\{\omega_i\}$ allow for assigning different weights to the k initial states, to encompass some possible differences in a population in empirical analyses; the shape of $v(|i-j|)$ assigns different degrees of importance to the intensity of movement between classes; the value of α defines a specific distance between P and the immobility matrix Id.

Besides meeting the axiomatic conditions of {N} and {M}, for the Alcalde-Unzu family the property of immobility is strong ({SI}) and two additional properties hold: Maximum Mobility ({SMM}) and Independence of Irrelevant Classes ({IIC}) (Alcalde-Unzu, Ezcurra, Pascual (2006) at pages 2 and 3). Evidently, the computation of this index is prohibitive, in case of big databases, without a suitable informatic device.

It is to be underlined in MarkovInfer_MobInd.m the indices of Ferretti-Ganugi and by Alcalde-Unzu are computed for the same set of weights ω_i, as well as for the same shape of the modulating function $v(\cdot)$, whatever the user choice. Moreover, *we chose to fix a not trivial* $\alpha = 1{,}3$.

8.5 Some not-directional mobility indices connected to the spectral structure of the transition matrix

The remaining indices, implemented in our Matlab tool and preceding by many years the ones by Alcalde-Unzu and by Ferretti-Ganugi, have nothing to do with mobility as an oriented direction or as a not-oriented movement, rather they are connected to the eigenspace's structure of the transition matrix P.

Among the long-standing mobility proxies, there are the Trace Index and the Determinant one (Shorrocks, 1978), respectively defined as:

$$I_{tr}(P) = \frac{k - tr(P)}{k-1} \quad \text{and} \quad I_{\det}(P) = 1 - |\det(P)| \qquad [8.9]$$

In virtue of its simplicity, $I_{tr}(P)$ has been appreciated until the beginning of the current Millennium; its continuous variant has been employed to sum up the

overall mobility inside samples inspected through Bayesian inference for Continuous-Time Markov Chains, relying on the intensity matrix (Fougère and Kamionka (2003)). $I_{\det}(P)$ is zero whenever the transition matrix exhibits at least one null row or one null column: it is a very common trait for transition matrices estimated over real economic data from large samples, thence it has been abandoned. Nevertheless, its extended version (here called as Exponentiated or Amplified or Generalized Determinant Index) in

$$I^{\beta}_{\det}(P; \mathrm{T}) = 1 - |\det(P)|^{\beta T} \qquad [8.10]$$

where β is a positive real number and T is the length of the temporal interval on which the transition matrix P is estimated, is one of the very few measures obeying to the Period Invariance ({PI}, Shorrocks (1978), page 1021). In the program we have fixed $\beta = 1{,}5$.

The Asymptotic-Half-Life Index $I_{\mathrm{H}}(P; \mathrm{T})$ is period-invariant, too (Shorrocks (1978), pages 1021-1022):

$$I_{\mathrm{H}}(P; \mathrm{T}) = \exp_e(-h\mathrm{T}) \qquad [8.11]$$

$$h \equiv -\log_e 2 \,/\, \log_e(|\lambda_2|)$$

The parameter h is the asymptotic half-life, indicating the speed of convergence towards the equilibrium distribution of a discrete-time, stationary, 1st-order Markov chain whose transition matrix is P; λ_2 is that eigenvalue inside the spectrum of P whose modulus is the second highest. $I_{\mathrm{H}}(P; \mathrm{T})$ is normalised into [0,1] because h ranges from 0 ($\lambda_2 = 0$) to infinity ($|\lambda_2| = 1$). The correspondence between mobility and convergence speed can be intuited from the fact that a rigid system with a slowly changing distribution, implying a slow convergence. On the other side a perfectly mobile population should reach the equilibrium distribution on a short time interval. When employing this index, it must be paid attention to the very fast convergence to zero of the negative exponential function on the positive real halfline.

We have interpreted as grounded on the transition matrix also two measures involving directly the eigenvalues of P: the Second-Eigenvalue Index $I_2(P)$ and the All-Eigenvalues Index $I_e(P)$ (Shorrocks (1978); Sommers and Conlisk (1979)), respectively:

$$I_2(P) = 1 - |\lambda_2| \quad [8.12]$$

$$I_e(P) = \frac{k - \sum_{i=1}^{k} |\lambda_i|}{k-1} \quad [8.13]$$

The last two mobility proxies in our device are the Index of predictability (Parker and Rougier (2001)):

$$I_{\text{predict}}(P) = \frac{k}{k-1}\left(\sum_{i,j=1}^{k} p_{ij}^{\,2} - 1\right) \quad [8.14]$$

and which is probably the second most famous index by its inventor, the Second Index of Bartholomew (Bartholomew, 1982):

$$I_{B,2}(P) = \frac{k}{k-1}\sum_{i,j=1}^{k} p_{ij}\,|i-j| \quad [8.15]$$

It is easy to notice a link with the Directional Index: indeed $I_{B,2}(P) = |I_{\omega,v}(P)|$ when $\omega_i \equiv 1/(k-1)\ \forall i=1, 2, ..., k$ and $v(|j-i|) \equiv |j-i|$.

8.6 Mobility of Italian companies according to transition-matrix-based indices for a 1st-order Markov chain

Regarding the quantitative analysis of the concept of mobility as inflected through the nine indices, computed by our program, from the great amount of elaborations spread over the 28 workspaces (Appendix A3), various regularities clearly emerge.

For the two simplest Markov frames, both in the year-by-year (or mono-period, or one-step) variant reported in the section A3.1, and in the aggregated-in-time (or compound, or multi-date, or multi-period, or multi-step) variant reported in the section A3.2, whatever the sector, two measures conceived by Shorrocks reveal a shortcoming: they steadily assume too extreme values, on the basis of our actual firms' data, to be useful and manageable, although they are in theory normalized into the real interval [0,1]. Indeed, Generalized Determinant Index remains too close to 1, while Asymptotic Half-life Index remains too close to 0. The flaw of the former comes from the fact that the transition matrix is

stochastic and, when estimated from actual economic data, is affected by the clustering around the main diagonal; in addition, we chose to split into ten sub-intervals the backing space of the Markov chains. The clustering around the main diagonal is the distribution of most of the probability mass above all on the principal diagonal and, secondarily, in the positions next to it. Thence, the transition matrix has positive components, lower than 1 and summing to 1 in every row; in case of real economic data, its structure is very similar to the one of identity matrix. Its determinant is approximated by the product of the sole entries inside the main diagonal and the more numerous are the Markov states, the closer the determinant is to 0: the associated index assumes a value very near to 1 when the states are many (e.g., the ten classes in our analysis). Instead, when the index is exactly 1, it occurs because the underlying matrix has at least one row entirely filled with zeros, namely one Markov stratum at least is empty at the beginning of the period. The latter index is one of the few time-invariant measures (invariance respect to time {PI} would ideally be a desirable property for a good mobility measure but it is very difficult to be rendered analytically) but is also defined by exponentiating the Euler's constant "*e*" (the negative exponential function converge fast to zero) by the opposite of the positive asymptotic half-life, whose value is generally very great when referring to a real economic system: thus, it is inevitable this index, associated to the half-life, goes quickly to zero.

On the other hand, the Index of Predictability and the 2nd Index of Bartholomew are not originally normalized into the conventional range [0,1], rather they are worth always visibly more than 1. For the one-step mobility, the former's values go approximately from 4.2 to 6.3 and the values of the latter approximately from 1.8 to 7 (obviously, for the compound multi-period mobility, they are even higher). The Trace Index and the All-Eigenvalues Index are the same measure when all the roots of the characteristic polynomial of the transition matrix are real numbers: it has been the case of our data.

It is possible to recognise a general ranking, valid for the four balance-sheet items of the "Five Traditional ATECO Manufactures", as well as of the InnoSMEs, for both the 1st order chains, for the mono-period (or one-step) mobility and for its multi-period (or multi-step or compound) extension. Numbers just below refer to the one-step mobilities. Index of Predictability assumes the greatest values, followed by the Second Bartholomew Index; then,

the Generalised Determinant measure (close to 1) and the All-Eigenvalues/Trace Index (approximately ranging from 0.17 to 0.55); the 2nd Eigenvalue Index rarely ranges between 0.25 and 0.16, more often it is worth less than 0.12. Finally, Index of Alcalde-Unzu seldom goes from 0.045 to 0.03 and very often stays under 0.022; it is very difficult to find the Directional Index by Ferretti and Ganugi between 0.023 and 0.012 and many of its estimations are significant only at the thousandth digit, when it is not negative. Surely, the last two measures could be manipulated to magnify variations by studying different inputs for the shared modulating function $v(|i-j|)$, the shared weights $\{\omega_l\}$ and for the Alcalde-Unzu power order α.

Let us focus on peculiarities of the Food & Beverages Industry, to provide an instance of analysis for the date-by-date indices. Here, monotonicity is sharp for Total Sales, it is not for the other variables. Index of Predictability is the unique increasing one but recall that, in truth, "predictability" is ideally opposite and complementary to "mobility"): sales' mobility inside Food manufactures tends to decrease during the decade 2006-2011. The 2nd Eigenvalue Index stands above the Alcalde-Unzu Index, diminishing quickly, until 2010-2011, and their estimations are nearby from 2011-2012. The Directional Index detects a prevailing weak sales' upsizing (i.e. it is positive, except for the span 2008-2009) from 2006 to 2011 and a net downsizing (i.e. it is always negative) from 2011 to 2015: in such a situation, its exclusive skill in highlighting an overall downwards dynamics is a strong advantage! Total Production is in practice equivalent to Total Sales for many Italian manufactures: alternation of signs of the Directional Index is the same for these two items, but all indices reveal a more pronounced decrease between 2006-2007 and 2007-2008 than for Total Sales. Total Assets seem to behave differently and, in general, its mobility is more fluctuating: for the step 2007-2008 all indices show a major peak, followed by a minor one at 2009-2010. The 2nd Eigenvalue Index is the unique to exhibit a fall on 2011-2012 for jumps among asset-values' intervals. Assets' mono-period Directional mobility is negative only between 2011 and 2013: the year-by-year trend during 2006-2015 for assets' endowment is upwards, also on 2013-2015, which may indicate an attempt throughout the decade to react to sales-and-production downsizing by investing. Total Debts is the balance-sheet item presenting on average the highest estimation for all mobility proxies. Debts patterns display

on the first half of the decade a more evident convexity than sales, production and assets, whose maxima at 2006-2007 and 2009-2010 flank the minimum between 2007 and 2009. After 2010 the Second Index of Bartholomew, the Directional one by Ferretti-Ganugi and the 2nd Eigenvalue mobility fluctuate whereas the others decrease. Precisely, the Directional mobility attests, from 2011 and 2015, a tendency to a debt downsizing hard to understand, owing to its simultaneity with downsizing respect to sales and production. Nevertheless, it must be noted that downsizing magnitude depends on time and vary alternately for every variable from a scale of $\sim 10^{-3}$ to a scale of $\sim 10^{-2}$. Finally, for annual mobility a break is observed at 2011: before 2011, upsizing and moderate flare-ups are important; afterwards, lowering and specific downsizing overcome.

Researchers, interested in socio-economic mobility, argue for the conjecture that a multiperiod measure of the mobility about any economic proxy should increase if the corresponding epoch expands (inflation according to the temporal spans' dilation: see the empirical studies on incomes in Atkinson, Bourguignon, Morrisson (1992)). Independently from the industrial sector and the variable, it is true for five of the nine indices, i.e. the Alcalde-Unzu family, the Trace Index and the All-Eigenvalues one, the 2nd Bartholomew Index and the 2nd Eigenvalue one. Such a conjecture is confirmed by the Index of Predictability, too (whose multi-period version always decreases over wider and wider epochs), bearing in mind that the concept of configuration's predictability of an evolving system may be interpreted as the opposite to the concept of mobility or, equivalently, mobility itself may be intended to be directly proportionate to the degree of unpredictability of the final configuration of the dynamical system. On the contrary, the compound-in-time Determinant Index, together with Asymptotic Half-Life one, tends to diminish; it is hard to explain because these measures are defined as concordant with the primary notion of mobility, not in opposition with it. The inflation conjecture can be true for the Directional Mobility, by Ganugi and Ferretti, exclusively if the date-by-date (or mono-period) values of the Index are constantly positive or constantly negative, during the whole epoch; it is a necessary consequence of its own definition which also includes the orientation, that is the sign, of translocations of individuals between strata. The common feature for positive indices, inflating while the time interval is expanding, is that such inflation was

fast for aggregated transitions in the first half of the decade, from 2006-2007 until 2006-2011, but it becomes slower and slower as time expansion has been advancing into the second half, from 2006-2012 to 2006-2015.

Time-aggregated measures through *MarkovInfer_MobInd.m* enable comparisons between mobility estimations subtended by the two simplest Markov chains: whatever the index, the general trend is that mobility on the not-stationary model is almost always higher than mobility determined by the stationary model, for the 'traditional ATECO Five' as well as for Inno-SMEs, with the unique exception of the SMEs' Total Debts. Notwithstanding, the gap between Markov stationary mobility and Markov not-stationary mobility lessens while the historical interval, starting at 2006, becomes broader and broader, until the overall decade 2006-2015 is drowned out: in this last case, multi-period mobility on the time-homogeneous chain exceeds the one on the not-stationary chain.

Focusing on the economic variables, from 2006 until 2015 Italian enterprises of the five ATECO division have generally displayed the highest mobility values relatively to Total Debts, the lowest ones relatively to Total Assets, both in the date-by-date sense and in the time-compound sense.

Whatever the proxy and periodization, in the same decade, the Machinery-and-Equipment Industry generally is the most mobile, followed by Rubber-and-Plastic Manufactures; then, Textile Industry is more mobile than Chemical Products, whose mobility is quite akin to the case of Food & Beverages.

Conclusion

From the half of the 1950s, discrete-time stationary first-order Markov Chains, subtended by a discrete-state space, began to be rigorously and systematically employed in Statistical Economics to describe the evolution of people from consecutive generations among social classes, of households among income levels, of enterprises among firm-size classes. Little by little some mathematical expressions of indices, involving transition probabilities estimated for the first-order stationary chain, were formulated. Then, the axiomatic theory of mobility indices, explicitly defined by functionals of the transition matrix, was conceived. Originally, Markov statistical analysis was applied at a national scale, availing of limited samples of microdata, consisting of few hundreds of economic agents, mimicking with difficulty the population to be studied (e.g., adult British employees, U.S. steel industry, farms or dairies of a specific State of the U.S. Federation etc.). While statistical inferential procedures (comprising hypothesis tests) were being developed not only for samples of granular data, but also for census data and for time series about Markov Chains, microdata sets were becoming larger and larger, thus more and more representative of the related population, as well as more and more demanding from a computational point of view. Academics stopped to take *a priori* for granted any reliability of Markov first-order stationarity when it began to be frequently rejected, relatively to the basic variables of Industrial Organization, against the second-order stationary chain and the Markov first-order time-inhomogeneity by the dedicated hypothesis tests, based on the Maximum Likelihood's Principle. Ergo, some statisticians interested in economics and social sciences turned their attention to aggregated data (cheaper, easy to manage and without any possibility of mathematical assessments on dependability). Some other ones afforded researches to more complex Markov models, exploiting mixture of diverse chains (like the mover-stayer model); or incorporating into the sample dynamics the entrances of newcomers and the departures of veterans respect to the market (by further probability distributions to make more sophisticated the 1^{st}-order transition matrix of the constant core of the sample and by introducing a reservoir as a

source of new members and as a repository for departed units); or indirectly inserting the dependence on time into the 1st-order chain via a direct dependence on economic determinants changing with the time, by an econometric approach. From the half of the 1980s, 1st-order stationary Markov Chains have been being also deployed onto auxiliary stochastic processes derived from time series, often representing tick-by-tick variations in the series itself, to highlight potential structural asymmetries, or asymmetries foremost due to the impact of economic cycles, in investments and unemployment rate.

We believe truth is intrinsic to microdata. Thus, in AIDA, the electronic archive of all Italian businesses, we assembled and exported 20 xlsx-type granular spreadsheets selecting the almost-complete populations of five traditional, historically relevant, Italian manufactures at the national scale respect to four balance-sheet variables: Total Assets, Total Sales, Total Debts, Total Production, all valued in euros. The five industries, classified by the ATECO 2007 system, are: Food and Beverage Industry (C10; ~13700), Rubber and Plastic Products (C22; ~7100), Textiles (C13; from 5700 to 6200), Machinery and Equipment Industry (C28; from 16900 to 18100), Chemicals (C20; ~4000), where in brackets the ATECO identity number is followed by the approximated number of companies. For each of the four variables, we have also built up the maximal sample of the very peculiar, low-sized sub-population of Innovative Small & Medium Enterprises (Innovative SMEs), which is registered in AIDA even it is outside the ATECO 2007 taxonomy. It is worth studying those latter ones because they are not absolutely-small-sized panels drawn from a much larger population, rather they must be interpreted as almost-complete samples, consisting of ~300 firms, from an 'elitist' and small business category. We have had at disposal the epoch from 2006 until 2015, marked by annual observations whatever the industry and whatever the variable: it is intriguing because in its middle there is the 2011 year, which is special as a 'shocking year' since in such a year Italy has begun to experience the delayed impact of the 2007-2008 international Great Recession. All 24 excel files are structured as bidimensional matrices, where every row refers to one, and only one, statistical unit (a precise firm) while every column bi-univocally corresponds to one, and only one, date of annual observation, when the final value of the economic variable was detected. It is an ideal data structure to be imported into the Matlab environment and to be processed by a

suitable Matlab code tailored for inferential estimation methodologies about discrete-time Markov Chains and transition-matrix-based mobility measures.
Our interactive device *MarkovInfer_MobInd.m*, expounded in the central part of this doctoral dissertation, is such a kind of code, through which we elaborated the 24 AIDA datasets creating as many corresponding Matlab workspaces full of intermediate and final outputs. We fashioned it to infer estimations some features of Markov Chains' theory according to raw data and to support the analyses in the second part of the thesis, whose outcomes are tabulated in a minimal part inside Appendix A2, Appendix A3 and in Chapter 7. *MarkovInfer_MobInd.m* (whose script has been integrally copied in Appendix A1) is capable to import and process automatically the microdata of a single scalar variable from an unbalanced wide-format longitudinal panel stored inside a xls/xlsx-type file. The user is requested to insert, one at a time into the prompt line of the Matlab Command Windows, all inputs needed for calculations. Among the inputs there are the name of the primary file exported from AIDA, the number of classes, and their extremes, splitting the variable's domain to construct the Markov state-space (over which the sample's individuals will be automatically distributed); the shapes or values of all parameters characterizing some of nine mobility indices designed as functionals of the transition probabilities by the scientific literature.
While inspecting original longitudinal spreadsheets of businesses, we reflected on the unbalancing problem, namely the fact that for many agents there were no observation for the variable at some dates: only a core of individuals was expressing a datum at every year during the decade (the so called incumbent or permanent units). Such dynamics of appearances and disappearances inside the spreadsheet is a more generic and unclean phenomenon than the genuine entry-exit dynamics of companies relative to the real market. It includes both mergers and acquisitions, as well as arrivals of new-born firms and true failures or liquidations of some veterans, and other spurious situations like communications flaws. Distinguishing the formers from the latter ones is a quite thankless job inside an industry consisting of thousands and thousands of companies. So, we decided not to implement any model of ingresses of actual market-newcomers' and of actual market-leavers' demises. Consequently, we opted for preliminarily extracting, by default, the generation of all incumbents during the entire registered epoch, from the exported longitudinal panel,

obtaining a balanced core to be processed. We judged it was the most tractable and the most influencing industrial sub-system to which pay attention; we have also no doubt the maximal sample of incumbents from AIDA is an almost-complete representation of the associated sub-population of residential enterprises, as in the case of the whole population, comprehending general appearances and disappearances, of each chosen manufacture. The residents during the decade 2006-2015 are inside the brackets of the following list: Food and Beverage Industry (~4500), Rubber and Plastic Products (~3300), Textiles (~2700), Machinery and Equipment Industry (~8200), Chemicals (~2000), Innovative SMEs (115). We found ratios of size of the incumbents' generation versus size of the overall panel range approximately from 30% to 50%; these percentages are strongly conditioned by the manufactural sector and weakly by the variable.

The 24 imported panels of permanent agents have been examined simultaneously through the Pearson's χ^2 test and the Maximum Likelihood Ratio Criterion (MLRC) for Markov Chains, executed by the code, concerning the same null hypothesis of first-order stationarity against the two distinct alternatives of first-order time-inhomogeneity (along with the null, it is *Query a*) and second-order stationarity (along with the null, it is *Query b*). We have indicated by the term 'antinomy', borrowed from philosophy, any couple of a null hypothesis compared with an alternative. In the time-homogeneous Markov scheme, for all the four variables of the five ATECO industries and the Inno-SMEs group, the first-order chain is defeated by the second-order chain due to infinitesimal *p*-values (there is a unique inconsistency at the 8,94% MLRC chance for Inno-SMEs Total Assets). Instead, apropos of the first-order Markov approach, the null of the stationary chain is generally rejected versus the not-stationary one for the five ATECO divisions, as proved by infinitesimal or decisively minor *p*-values; nevertheless, the stationary first-order becomes more credible in the case of Innovative SMEs, where occurrence probabilities for this null are above the 60% in case of Total Sales and Total Debts, and between 25% and 55% for Total Assets (however, admissibility percentages of 5,40% and 7,92% for Total Production cannot to be completely discarded in comparison with the usual threshold of 5%).

As for hypothesis tests employing the chi-squared distribution, some economists are beginning to belief that χ^2-type tests are inclined to reject the null in case of big data. For us if such a test refuses the null relatively to a dataset, then such a null is inappropriate for the current set. Rather, what matters is how much representative the panel is compared to the population holding it, when a null hypothesis is overruled for the panel by the Pearson's test or by the MLRC. Moreover, potential χ^2-type-test's tendency to refuse the null may have not a logical sense for Markov Chains' features. Indeed, chi-squared statistics of each of the two antinomies is the ratio between the maximised likelihood functions of the current null and of its alternative. When swapping the roles of the null and of its alternative, also the ratio between the maximised likelihood functions must be inverted: the same statement would be overruled in the role of the null before the swap, in the role of alternative after the swap. A fairer discussion point is sensibility to little variations in very large samples. The χ^2 little-deviation-sensitivity normally emerges in big samples of many tens of thousands, or in some hundreds of thousands, of individuals. Instead, sizes of our panels from manufactural ATECO divisions, as well as the size of the InnoSMEs group at a greater extent, are only some thousands, so we are sure little-deviation-sensitivity has affected very modestly our χ^2-tests' p-values, which are almost equal to zero in disfavour of Markov first-order stationarity.

Therefore, we are convinced that respect to a national sample (i.e. filled with thousands of units) of a certain Italian manufacture, highly representative of the overall actual population for the industry, data themselves claim an upper level of complexity in the Markov model than the first-order time-homogeneous chain. Such a major complexity may be achieved either implementing, into the Markov model, the historical path of the system via a higher-order chain (from the second to a greater one) or attaching a temporal pointwise dependence (i.e. not-stationarity) to the first order chain. We have conjectured that in each wide manufacture strong correlations may exist between businesses: horizontal correlations, in virtue of competition among subjects of comparable sizes, and vertical correlations, in virtue of relationships between subjects of different sizes, for instance of the kind customer-subcontractor. They are best captured through history dependence, namely the second or a higher chain order, or through the pointwise temporal dependence. On the contrary, first-order

stationarity seems to be dependable for a representative panel of a very strictly-determined and less large (some hundreds of units) sub-population of enterprises at the national scale, like the Innovative Small and Medium ones, outside the ATECO system, where inner correlations are likely very weak, at least. It happens as if this subpopulation is "diluted" within the Italian economic scenario, and such a condition may weaken horizontal correlations; while propensity to innovation may contribute to a sort of isolation and to direct connections with niche markets, preventing vertical correlations.

However, 2nd-order stationary chain is more appropriate according to both antinomies (*Query a* and *Query b*), for all four variables, for the five traditional ATECO sectors and the Inno-SMEs subpopulation. It suggests that history of the national manufactural system affects its evolution, regardless of the presence or the absence of correlations, but history dependence is a feature to exploit with care: the higher the order of the Markov chain is, the minor the number of sample's individuals to be involved into equations of inferential estimation's methods.

Concerning the analysis of the concept of mobility as inflected in a rigidly intragenerational sense through the nine indices in our program, various regularities clearly have emerged from elaborations spread over the 28 workspaces. Here are some examples.

For the two 1st-order chains, both in the year-by-year version and in the aggregated-in-time variant, whatever the sector, two Shorrocks's measures have a flaw: they really assume too extreme values to be helpful, although in theory normalized into the real [0,1]. The Generalized Determinant Index remains too close to 1, owing to the clustering on the main diagonal combined with the relatively great number of states; while Asymptotic Half-life Index remains too close to 0.

On the other hand, the Index of Predictability and the 2nd Index of Bartholomew are not normalized into the conventional range [0,1] and are always estimated more than 1. The Trace Index and the All-Eigenvalues Index are the same measure because all the roots of the characteristic polynomial of our transition matrices are real.

A general ranking can be identified for the four balance-sheet items of the "Five Traditional ATECO Manufactures", as well as of the InnoSMEs, for both the 1st -order chains, for the mono-period mobility and for its multi-period

extension. Index of Predictability assumes the greatest values, followed by the Second Bartholomew Index; then, the Generalised Determinant measure (close to 1), the Trace/All-Eigenvalues one and the 2nd Eigenvalue Index; finally, the measure of Alcalde-Unzu followed by the Directional Index

Focusing on Food & Beverages Industry, mobility is monotone for Total Sales, not for the other variables. Index of Predictability is the unique which increases: sales' mobility inside Food manufactures tends to decrease during the decade 2006-2011. The Directional Index reveals a weak sales' upsizing from 2006 to 2011 and a net downsizing from 2011 to 2015. For Total Production Directional Index's alternation of signs is the same for sales, but all indices reveal a more pronounced decrease between 2006-2007 and 2007-2008. Total Assets seem to behave differently and, in general, its mobility is more fluctuating. Assets' mono-period Directional mobility is negative only between 2011 and 2013: the year-by-year trend during 2006-2015 for assets' endowment is upwards, indicating an attempt to react to sales-and-production downsizing by investing. Total Debts presents on average the highest estimation for all mobility proxies. Debts patterns display on the first half of the decade a more evident convexity than sales, production and assets. The Directional mobility attests, from 2011 and 2015, a tendency to a debt downsizing hard to understand. Finally, for annual mobility a break is observed at 2011: before 2011, increasing mobility and upsizing are important; afterwards, lowering and specific downsizing overcome.

Concerning inflation according to the temporal spans' dilation, independently from the industrial sector and the variable, it is confirmed for five aggregated-in-time indices: Alcalde-Unzu family, the Trace/All-Eigenvalues one, the 2nd Bartholomew Index, the 2nd Eigenvalue one and, complementarily, by the Index of Predictability. The compound-in-time Determinant Index, together with Asymptotic Half-Life one, tends to diminish. The inflation conjecture can be true for the Directional Mobility if the mono-period values of the Index are constantly positive or constantly negative, during the decade. The shared connotate for the positive indices is that inflation was fast for aggregated transitions in the first half of the decade, from 2006-2007 until 2006-2011, but becomes slower while advancing into the second half, from 2006-2012 to 2006-2015.

Whatever the index, mobility on the not-stationary model is almost always higher than mobility determined by the stationary model, for the 'traditional ATECO Five' as well as for Inno-SMEs, with the unique exception of the SMEs' Total Debts. The gap between Markov stationary mobility and Markov not-stationary mobility lessens while the historical interval, from 2006, becomes wider and wider, until the whole 2006-2015, when multi-period mobility on the time-homogeneous chain exceeds the one on the not-stationary chain.

Overviewing on economic variables, from 2006 until 2015 Italian enterprises of the "Five ATECO Traditional Ones" have displayed the highest mobility values for Total Debts, the lowest for Total Assets, both in the date-by-date sense and in the time-compound sense. Whatever the index and periodization, the Machinery-and-Equipment Industry generally is the most mobile, followed by Rubber-and-Plastic Manufactures and Textile Industry; Chemical Productions' mobility is almost the same of Food & Beverages.

Bibliography

Alcalde-Unzu J., Ezcurra R., Pascual P. (2006) Mobility as movement: A measuring proposal based on transition matrices, *Economics Bulletin*, Vol. 4, No. 22, pp. 1-12.

Alexander W. H. (1965) Growth patterns and survival tendencies of firms in the Louisiana dairy industry, *LSU Agricultural Experiment Station Reports No. 203*, pp. 1-33 (http://digitalcommons.lsu.edu/agexp/203) and *Department of Agricultural Economics and Agribusiness Bulletin No. 593*, pp. 1-33, Agriculture Experiment Station, Louisiana State University.

Anderson T. W. (1954) Probability models for analysing time changes in attitudes, in P.F. Lazarsfeld (Ed.) *Mathematical Thinking in the Social Sciences*, The Free Press, Glencoe (IL).

Anderson T. W. and Goodman L. A. (1957) Statistical Inference about Markov chains, *The Annals of Mathematical Statistics*, Vol. 28, No. 1, pp. 89-110.

Adelman I. G. (1958) A Stochastic Analysis of the Size Distribution of Firms, *Journal of the American Statistical Association*, Vol. 53, No. 284, pp. 893-904.

Atkinson A. B., Bourguignon F., Morrisson C. (1992) *Empirical Studies of Earnings Mobility*, Harwood Academic Publishers, London (GB).

Bartholomew D. J. *Stochastic Models for Social Processes*, Wiley, London (editions from 1967 to 1982).

Bergh D. (2015) Chi-Squared Test of Fit and Sample Size - A Comparison between a Random Sample Approach and a Chi-Square Value Adjustment Method, *Journal of Applied Measurement*, Vol. 16, No. 2, pp. 204–217.

Bertoni D., Aletti G., Ferrandi G., Micheletti A., Cavicchioli D., Pretolani R. (2018) (equivalently Bertoni D., Aletti G. et Alii (2018)), Farmland Use Transitions After the CAP Greening: a Preliminary Analysis Using Markov Chains Approach, *Land Use Policy*, Vol. 79, pp. 789-800.

Bibby J. (1975) Methods of Measuring Mobility, *Quality and Quantity*, Vol. 9, No. 2, pp. 107-136.

Bickenbach F. and Bode E. (2003) Evaluating the Markov Property in Studies of Economic Convergence, *International Regional Science Review*, Vol. 26, No. 3, pp. 363–392.

Boudon R. (1973) *Mathematical Structures of Social Mobility*, Elsevier, Amsterdam.

Bourguignon F. and Morrison C. (2002) Inequality among World Citizens, *American Economic Review*, Vol. 92, No. 4, pp. 727-744.

Bostwick D. (1962) Yield Probabilities as a Markov Process, *Agricultural Economics Research*, Vol. 14, No. 2, pp. 49-56.

Champernowne D. G. (1953) A Model of Income Distribution, *The Economic Journal*, Vol. 63, No. 250, pp. 318-351.

Champernowne D. G. (1973) *The Distribution of Income between Persons*, Cambridge University Press.

Chavas J. P. and Magand G. (1988) A dynamic analysis of the size distribution of firms: the case of the US dairy industry, *Agribusiness*, Vol. 4, No. 4, pp. 315-329.

Ching W. K., Wang X., Ng M. K., Siu T. K. (2013) *Markov Chains - Models, Algorithms and Applications* (*2nd Ed.*), International Series in Operations Research & Management Science, Vol. 189, Springer.

Collins N. R. and Preston L. E. (1961, a) The Structure of Food-Processing Industries 1935-55, The Journal of Industrial Economics, Vol. 9, No. 3, pp. 265-279.

Collins N. R. and Preston L. E. (1961, b) The Size Structure of the Largest Industrial Firms 1909-1958, The American Economic Review, Vol. 51, No. 5, pp. 986-1011.

Disney W. T., Duffy P. A., Hardy W. E. (1988), or Disney W. T. et Alii (1988), A Markov chain analysis of pork farm size distributions in the South, *Southern Journal of Agricultural Economics*, Vol. 20, No. 2, pp. 57-64.

Duncan G. T. and Lin L. G. (1972) Inference for Markov Chains Having Stochastic Entry and Exit, *Journal of the American Statistical Association*, Vol. 67, No. 340, pp. 761-767.

Dunne T., Roberts M. J., Samuelson L. (1988) Patterns of Firm Entry and Exit in U.S. Manufacturing Industries, *The RAND Journal of Economics*, Vol. 19, No. 4, pp. 495-515.

Ethridge. D. E., Roy S. K., Myers D. W. (1985), or Ethridge et Alii (1985), A Markov Chain Analysis of Structural Changes in the Texas High Plains Cotton Ginning Industry, *Southern Journal of Agricultural Economics*, Vol. 17, No. 2, pp. 1-10.

Feller W. (1950) *An Introduction to Probability Theory and its Applications*, Wiley for New York-Chapman & Hall for London, Various Editions.

Ferretti C. (2012) Mobility as Prevailing Direction for Transition Matrices, *Statistic and Application*, Vol. 10, No. 1, pp. 67-78.

Ferretti C. and Ganugi P. (2013) A new mobility index for transition matrices, *Statistical Methods & Applications*, Vol.22, No. 3, pp. 403-425.

Fields G. S. (2000) Income mobility: Concepts and measures-Patterns and underlying causes, in N. Birdsall & C. Graham (Eds.) *New markets, new opportunities? Economic and social mobility in a changing world*, pp. 101-132, The Brookings Institution Press, Washington (DC).

Fields G. S. (2008) Income mobility, in L. Blume & S. Durlauf (Eds.) *The new Palgrave dictionary of economics*, Palgrave Macmillan, New York (NY).

Fields G. S. and Ok E. A. (1996) The Meaning and Measurement of Income Mobility, *Journal of Economic Theory*, Vol. 71, No. 2, pp. 349-377.

Fields G. S. and Ok E. A. (1999) The measurement of income mobility: An introduction to the literature, in J. Silber (Ed.) *Handbook on income inequality measurement*, pp. 557-596, Kluwer Academic Publishers, Norwell (MA).

Fingleton B. (1997) Specification and testing of Markov chain models: An application to convergence in the European Union, *Oxford Bulletin of Economics and Statistics*, Vol. 59, No. 3, pp. 385-403.

Fougère D. and Kamionka T. (2003) Bayesian inference for the Mover-Stayer model in continuous time with an application to labour market transition data, *Journal of Applied Econometrics*, Vol. 18, pp. 697-723

Glass D. V. (1954), Editor, *Social Mobility in Britain*, Routledge & Kegan Paul.

Gort M. and Klepper S. (1982) Time Paths in the Diffusion of Product Innovations, *Economic Journal*, Vol. 92, pp. 630-653.

Grimmett G. R. and Stirzaker D. R. (2001) *Probability and Random Processes*, Oxford University Press, New York (NY), 3rd Edition.

Hallberg M. G. (1969) Projecting the Size Distribution of Agricultural Firms - An Application of a Markov Process with Non-Stationary Transition Probabilities, *American Journal of Agricultural Economics*, Vol. 52, No. 2, pp. 289-302.

Hart P. E. and Prais S. J. (1956) The analysis of business concentration: a statistical approach, *Journal of the Royal Statistical Society (Series A)*, Vol. 119, No. 2, pp. 150-191.

Horowitz A. and Horowitz I. (1968) Entropy, Markov Processes and Competition in the Brewing Industry, *The Journal of Industrial Economics*, Vol. 16, No. 3, pp. 196-211.

Judge G. G. and Swanson E. R. (1962) Markov Chains: Basic Concepts and Suggested Uses in Agricultural Economics, *The Australian Journal of Agricultural and Resource Economics*, Vol. 6, No. 2, pp. 49-61.

Klepper S. and Graddy E. (1990), The Evolution of New Industries and the Determinants of Market Structure, *The RAND Journal of Economics*, Vol. 21, No. 1, pp. 27-44.

Krenz R. D. (1964) Projections of farm numbers for North Dakota with Markov chains, *Agricultural Economics Research*, Vol. 16, No. 3, pp. 77-83.

Lee T. C., Judge G. G., Takayama T. (1965) On Estimating the Transition Probabilities of a Markov Process, *Journal of Farm Economics*, Vol. 47, No. 3, pp. 742-762.

Lee T. C., Judge G. G., Zellner A. (1968) Maximum Likelihood and Bayesian Estimation of Transition Probabilities, *Journal of the American Statistical Association*, Vol. 63, No. 324, pp. 1162-1179.

Martinez A. R. and Martinez W. L. (2007) *Computational Statistics Handbook with MATLAB, 2nd ed.*, Chapman & Hall/CRC.

Mellor C. J. (1984) An application and extension of the Markov chain model to cereal production, *Journal of Agricultural Economics*, Vol. 35, No. 2, pp. 203-217.

Neftci S. N. (1984) Are Economic Time Series Asymmetric over the Business Cycle? *Journal of Political Economy*, Vol. 92, No. 2, pp. 307-328.

Nelson C. H., Braden J. B., Roh, J. S. (1989) Asset fixity and investment asymmetry in agriculture, *American Journal of Agricultural Economics*, Vol. 71, No. 4, pp. 970-979.

Padberg D. I. (1962) The Use of Markov Processes in Measuring Changes in Market Structure, *Journal of Farm Economics*, Vol. 44, No. 1, pp. 189-199.

Palm W. J. (2010) *Introduction to MATLAB for Engineers*, *various editions*. McGraw-Hill Higher Education.

Papoulis A. and Pillai S. U. (2002) *Probability, Random Variables and Stochastic Processes*, McGraw-Hill Higher Education, New York (NY), 4th Edition.

Pareto V. (1897) Cours d'économie politique, *Various Editions*.

Parker S. C. and Rougier J. (2001) Measuring Social Mobility as Unpredictability, *Economica*, Vol. 68, No. 269, pp. 63-76.

Piet L. (2008) The evolution of farm size distribution: revisiting the Markov chain model, *12th International Congress of the European Association of Agricultural Economists* (EAAE 2008 on August 26-29 in Ghent, Belgium), Record No. 44269, pp.1-10.

Prais S. J. (1955) Measuring Social Mobility, *Journal of the Royal Statistical Society (Series A)*, Vol. 118, No. 1, pp. 56-66.

Schluter L. (1997) On the Non-Stationarity of German Income Mobility, *LSE Distributional Analysis Research Programme, Discussion Paper No. 30*. London School of Economics, London.

Shorrocks A. F. (1976) Income Mobility and the Markov Assumption, *The Economic Journal*, Vol. 86, No. 343, pp. 566-578.

Shorrocks A. F. (1978) The Measurement of Mobility, *Econometrica*, Vol. 46, No. 5, pp. 1013-1024.

Sommers P. and Conlinsk J. (1979) Eigenvalue Immobility Measure for Markov Chains, *Journal of Mathematical Sociology*, Vol. 6, pp. 253-276.

Stanton B. F. and Kettunen L. (1967) Potential Entrants and Projections in Markov Process Analysis, *Journal of Farm Economics*, Vol. 49, No. 3, pp. 633-642.

Stavins R. N. and Stanton B. F. (1980) Using Markov Models to Predict the Size Distribution of Dairy Farms, New York State, 1968-1985, *Agricultural Economics Research Bulletin No. 181522 of the Cornell University Agricultural Experiment Station* (edited by C. H. Dyson School of Applied Economics and Management), Vol. 80, No. 20, pp.1-49.

Sutton J. (1997) Gibrat's Legacy, *Journal of Economic Literature*, Vol. 35, No. 1, pp. 40-59.

Tan B. and Yilmaz K. (2002) Markov chain test for time dependence and homogeneity: An analytical and empirical evaluation, *European Journal of Operational Research*, Vol. 137, No. 3, pp. 524-543.

Telser L. G. (1962) The Demand for Branded Goods as Estimated from Consumer Panel Data, *Review of Economics and Statistics*, Vol. 44, No. 3, pp. 300-324.

Williams D. C. (1963) A stochastic analysis of size distribution of firms in fluid milk markets in Louisiana, *LSU Agricultural Experiment Station Reports No. 411*, pp. 1-64 (https://digitalcommons.lsu.edu/agexp/411) and *Department of Agricultural Economics and Agribusiness Bulletin No. 578*, pp.1-64, Agriculture Experiment Station, Louisiana State University.

Zimmermann A., Heckelei T., Domínguez I. P. (2009) (equivalently Zimmermann A. et Alii (2009)), Modelling farm structural change for integrated ex-ante assessment: Review of methods and determinants, *Environmental Science & Policy*, Vol. 12, No. 5, pp. 601-618.

Webography

[1] www.bvdinfo.com/en-gb/our-products/data/national/aida

[2] www.infocamere.it/home

[3] www.istat.it/en/tools/glossaries-and-classifications/migration-to-ateco-2007

[4] www.istat.it/it/strumenti/definizioni-e-classificazioni/ateco-2007#codesearch

[5] www.registroimprese.it/l-anagrafe-nazionale-delle-imprese