Universiteit Utrecht

Innovation Studies Utrecht (ISU) Working Paper Series

UNRAVELLING START-UP PROCESSES WITH THE HELP OF SEQUENCE ANALYSES

Andrea Herrmann
Kim van der Putten

ISU Working Paper #14.02

UNRAVELLING START-UP PROCESSES WITH THE HELP OF SEQUENCE ANALYSES

Andrea M. Herrmann, Utrecht University, the Netherlands
Kim van der Putten, Wageningen University, the Netherlands

ABSTRACT

Our sequence analyses of the PSED2 database, the largest available dataset on venture creation processes, demonstrate two points. First, they show when and how to use the numerous variants of this method. To this end, we develop a decision tree that makes the analytical choices to be taken explicit. Since researchers can often not know *ex ante* which way of running sequence analyses delivers the most insightful results, we suggest to use this decision tree and: `in case of doubt, do it both ways!`. Second, our analyses also highlight the usefulness of sequence techniques for studying venture creation processes. Contrary to previous start-up analyses, we succeed in identifying 16 distinct ways in which entrepreneurs set up new ventures. These findings suggest that previous studies fail to recognize systematic venture creation patterns due to the use of traditional statistical techniques which do not make it possible to treat one sequence of events as one single case.

1. INTRODUCTION: SEQUENCE ANALYSES – A USEFUL METHOD FOR THE ANALYSIS OF SOCIAL SCIENCE DATA?

Researchers used to conducting empirical research are likely to agree that their findings depend on the method used to analyze data. Just think of a dataset on corporate performance, where static regression analyses are likely to deliver significantly different explanations of corporate success than dynamic time-series cross-sectional analyses. Over the past decades, the explosion of new information technologies has led to a massive increase in software and statistical tools for the analysis of empirical data. While the opportunities of making sense of empirical data have increased exponentially as a result, the risk of running non-sense analyses has increased alike. This paper is about the sense and non-sense or, rather, the use and usefulness of one of these new statistical tools for the analyses of social-science data: sequence analyses (SA).

Originally developed as a tool to decode the human genome, SA are increasingly used in the social sciences since the early 1980s (see Sankoff and Kruskal 1983). Contrary to traditional statistical methods, like time-series analyses, which treat one event sequence as several, stochastically generated data points, SA treat one sequence of events as one unit of analysis. More concretely, SA measure the 'distance' between one sequence and another by calculating the minimum number of sub-sequences that need to be changed in order to convert one sequence into another. By extending this logic to the entire sample, the most common processes can be identified whereby the type, order, and timing of events are discerned.

While SA have been successfully applied in natural science research, their usefulness for the analysis of social science data is still contested. Interestingly, this discussion seems to be motivated by the (confusingly) high number of decisions that researchers need to make in order to run SA. Certainly, SA are not easy to use, yet promise to deliver novel insights. To illustrate the use and

usefulness of SA, we study one of those topics that is of major interest to researchers across the social sciences: entrepreneurship. Since entrepreneurship has been recognized as a central driver of economic growth and, hence, prosperity (European Commission 1999; European Commission 2003; Audretsch and Thurik 2001; Carree and Thurik 2003; Stel, Carree et al. 2005), scholars in disciplines as diverse as management and business, sociology, economics, psychology, and the political sciences have been interested in the triggers, drivers, and success factors of entrepreneurship. Following the path-breaking argument of Gartner (1988) and Low& Macmillan (1988) that entrepreneurship ought to be understood as a process rather than a static event, fundamental questions that have gained momentum over the past two decades concern possible commonalities in venture creation processes (see, for example, Sarasvathy 2001; Ucbasaran, Westhead et al. 2001; Davidsson 2008). How do entrepreneurs proceed to start a new business? Do entrepreneurs with the same characteristics follow a similar path to set up a new company? Is it possible to discern how founding processes differ as a function of the entrepreneur`s start-up motive, his collaborators, geographic location or industrial sector? Seeking to answer these questions, more than hundred researchers at 36 US universities joined their forces to collect data for one of the most comprehensive databases on venture creation processes: the Panel Study of Entrepreneurial Dynamics (PSED).

Interestingly, analysts of the PSED data are rather sceptical whether and, if so, which ideal-typical processes of venture creation exist (Gartner, Carter et al. 2004; Hills and Singh 2004; Reynolds and Curtin 2008; see also Vesper 1980; Reynolds and Miller 1992; Cheng and Van de Ven 1996). Importantly, though, this conclusion is usually based on the use of traditional statistical tools. By focusing on venture creation processes at different points in time, the current literature thus fails to analyze one founding process as one single case. Accordingly, the order and duration (timing) of founding steps are not taken into account. Examples of these studies are numerous and include, most prominently, the work of Gatewood et al. (1995), as well as Carter et al. (1994; 1996) and, more recently, Gartner et al. (Gartner, Shaver et al. 2004), Reynolds & Curtin (2008), and Whittaker (2009).

Contrary to previous studies, we here use SA to analyze the PSED data with the aim of uncovering similarities in venture creation processes. Our results provide two major insights. First, we illustrate the usefulness of SA for the analysis of social science data as we identify 16 distinct ways in which entrepreneurs found new ventures. This finding suggests that previous studies might fail to recognize common founding sequences due to the use of traditional statistical methods. Second, we show how to use SA, by making its various steps as well as the choices related to each step explicit. This leads us to develop a decision tree which, we hope, will be of help to other researchers who consider using SA.

To illustrate these arguments, the paper is organized as follows. Section 2 sketches how SA works and gives an overview of the criticism raised against SA in the social science literature. Section 3 introduces various `hypotheses` on how to conduct SA. By discussing the various alternatives of running SA, the section gradually develops the decision tree presented at the end. Section 4 uses this tree to analyze the PSED dataset. Interestingly, these analyses show that there is, indeed, one particularly meaningful way of conducting SA. Applying this SA procedure to the PSED dataset, we identify 16 ideal-typical sequences along which entrepreneurs create new ventures. Section 5 summarizes and concludes the article by highlighting both the use and usefulness of SA for the assessment of social science data.

2. THE LITERATURES ON SA AND ITS CRITICS: HOW DOES SEQUENCE ANALYSIS WORK?

The use of sequence analyses for the assessment of social science data was first proposed by Andrew Abbott in the late 1980 (Abbott and Forrest 1986). Originally developed for the analysis of protein and DNA sequences of the human genome, the method initially received mixed reactions from social scientists. More recently, however, SA have been used with increasing frequency in social science research to describe life trajectories (Abbott and Tsay 2000; Wu 2000). While this increase in the application of SA is partly a result of ever more sophisticated statistical software, the question arises if, and - if so - under which conditions SA are useful tools for assessing social science data.

This first entails the question how SA work. What is so unique about SA? Possibly the most distinctive feature of SA which also differentiates it from other longitudinal research methods is its capacity to calculate `distances` between sequences. A 'distance' is the degree of similarity between two event sequences and is calculated with a specific computer algorithm. It is a further distinct characteristic of SA that the identification of similar sequences is usually not an end in itself, but typically combined with cluster analyses (see Kaufman and Rousseeuw 1990). Finally, the identification of similar sequences through cluster analyses is most meaningful if it is complemented with further statistical methods that explain or describe the sequence patterns obtained. Consequently, SA are typically carried out in four steps: 1) Events or activities are selected and ordered chronologically (building sequences), 2) The distance between all sequences is calculated and minimized (distance calculation), 3) the individual sequences are grouped according to their similarity (clustering, including cluster formation and cluster validation). 4) Finally, the clustered sequences are `explained` with the use of conventional statistical measures (e.g. cross-tabulations) or `described` through various graphical representations.

Let us turn to a concrete example which illustrates these steps. As mentioned in the introduction to this paper, entrepreneurship research is of interest to many social science disciplines and therefore provides a particularly comprehensive case for illustrating the applicability of SA. Remember that the **first step** consists in building sequences, whereby individual events, their order, and duration are recorded. For the sake of simplicity, just imagine that an entrepreneur, who wishes to open a new company, needs to undertake three steps: He needs to gather information about the market potential of his product; this step shall here be abbreviated with the letter `I`. He also needs to hire employees (abbreviated with an `E`) and he needs to acquire venture capital (short `V`). Accordingly, the sequence `I-E-V` describes a venture creation process where the entrepreneur first gathered information, then hired employees, and finally acquired venture capital. Contrary to that, an entrepreneur with the founding sequence 'I-V-E' also gathered information first, but then acquired capital before he hired collaborators. Importantly, sequences can not only capture the type and order, but also the duration (timing) of events. Accordingly, the founding sequence `V-I-E-E` does not only tell us that an entrepreneur first acquired capital, then gathered information, and finally hired employees. It also tells us that the last step took two units of time (maybe because qualified collaborators were particularly difficult to find.)

In the **second step**, the distances between all individual sequences are calculated. Remember that a 'distance' is the degree of similarity between two event sequences. More specifically, this

distance is defined as the number of transformations that are required to convert one sequence into the other. Importantly, each sequence is characterized by two aspects: *states* (in our example, the different types of entrepreneurial activities leading to venture creation) and the *order* of these states (Elzinga 2003). For the sake of clarity, note that a *state* has no length. An *event*, which we also call an activity or step, is a state with a length of one or several time periods. A *sequence*, in turn, is composed of several events. Sequences can be converted into one another gradually: either through so-called `substitutions` or through `insertions and deletions` (short `indels`). A substitution (also called replacement) is one single operation: one state is substituted with another state at the same point in time. An `indel` can consist of two operations: first a new state is inserted at one point of the event chain and then another state is deleted at an earlier or later point of the chain (insertion and deletion). Importantly, insertions and deletions can be carried out individually, i.e. a new state can be inserted while no other state is deleted (insertion), or a new state is deleted while no other state is inserted (deletion). Note that single insertion or deletion operations imply that events and, hence, sequences change in length. Note also that the substitution of states implies that the initial order of states is preserved while the states` types are altered. Contrary to that, the subsequent insertion and deletion of steps implies that their original order is changed whereas their original types are preserved (Lesnard 2006). It is necessary to understand these distinctions, because various types of SA differ in how their algorithms use substitution or indel operations respectively to calculate distances (see section 3).

Importantly, a researcher must assign `costs` for both substitution as well as indel operations. With regard to substitutions, costs are listed in a so-called substitution matrix. This table lists and assigns values (`costs`) to all possible combinations of substitutions of states.[1] Substitution (or replacement) costs can be assigned in three ways: a *fixed* hierarchy or classification can be compiled that ranks individual substitutions and assigns costs accordingly. Furthermore, replacement costs can be assigned *manually* based on the researcher`s experience with the occurrence of various states (educated guess), or *automatically* with the use of a transition matrix (MacIndoe and Abbott 2004). In the latter case, a table is built that reports how often certain types of states follow each other. Assuming that frequent subsequent states are often more strongly related, transition rates can be used to estimate the replacement costs. With regard to indel operations, their costs are to be calculated in relation to substitution costs. This procedure is known as the `indel cost` and for mathematical reasons it is generally a fixed number. To conclude, note that the costs assigned to substitution and indel operations respectively determine whether greater emphasis is given to preserving the order or the types of states. Section 3 briefly recapitulates how different algorithms set substitution and indel costs respectively.

Let us return to our entrepreneurship example for an illustration of distance calculation. Let us assume that all individual events are equally important for venture creation so that the substitution of one by another state is equally costly. The cost for one substitution shall be equal to 1 and the cost of one insertion or deletion shall be 0.5, so that one complete indel operation costs 1 point. To transform `I-E-V` into 'I-V-E', a researcher can thus choose different approaches. He can, for

[1] Note that different costs can be assigned to different `directions` of replacement. In other words, a researcher can make it more costly to replace A by B than to replace B by A. In the social sciences, this opportunity is important because different events do often not have the same meaning. Supposed that entrepreneurs find it generally more difficult to acquire venture capital (V) than hire employees (E), the relative importance of V and E for venture creation would be reflected if the substitution of V by E is less costly than the replacement of E by V.

example, only use substitutions and first substitute the `E` state of the first sequence by a `V`. He can then substitute `V` by an `E`. In this way, `I-E-V` is turned into 'I-V-E' for the cost of 2. Alternatively, a researcher can perform one complete indel operation: insert a `V` after the `I` of the first sequence and then eliminate the `V` from the end of this sequence. This transformation would have a cost of 1. Note that computer algorithms are programmed to identify the smallest difference between two sequences, i.e. the least costly transformations. Importantly, though, algorithms differ in their use of substitution and indel operations. Some algorithms only use substitution but no indel operations. Supposed an algorithm only uses substitution operations, it would choose the first aforementioned transformation approach. An algorithm that can perform both substitution and indel operations would choose the second approach. By comparing any of the venture creation sequences to each other, a computer algorithm identifies the minimum number of substitution (and indel) operations needed to convert one sequence into the others. Thereby, the most common sequence can be identified. Akin to a regression line, this most common sequence is the one that, overall, requires the lowest number of operations to be transformed into any of the sequences observed. While the importance of transformation costs for SA will become clear from our analyses of the PSED dataset in section 4, this short example illustrates that transformation costs and, hence, the results of SA, depend a) on the transformation operations used and b) on how the costs of these operations are set.

Let us turn to **step 3**. Thus far, we saw that SA distils the most common sequence from a broader set of event sequences. Note, however, that it is often not particularly meaningful to identify just one sequence, even if the latter is most representative of all observations made. The reason is straight-forward. Often several groups exist, because some bundles of sequences are more similar to each other than other bundles. Returning to our entrepreneurship example, we see that the sequences `I-E-V` (no. 1) and 'I-V-E' (no. 2) are more similar to each other. The same holds for the sequences `V-I-E-E` (no. 3) and `V-I-I-E (no. 4)`. However, both sequences 1 and 2 are comparatively different from sequences 3 and 4 respectively. To detect such groups of most similar sequences, cluster analyses are particularly useful: they assess event patterns and identify the optimum number of groups for which similarities are maximized within and minimized between groups. Since this paper is about SA rather than cluster analyses, we shall not discuss cluster algorithms in detail here. While the opportunities related to different types of cluster analyses are discussed together with our analyses of the PSED dataset (sections 3 and 4), suffice it to say here that SA are usually conducted in combination with cluster analyses to uncover the most meaningful number of similar sequences leading to an outcome.

In a fourth step, researchers typically want to understand what causes the differences between groups of most similar sequences. In other words, which factors cause events to unfold in one rather than another way? Let us assume that cluster analyses of our entrepreneurship data found two groups of most common sequences: `I-V-E` (group 1) and `V-I-E-E` (group 2). Under which conditions are entrepreneurs more likely to open a firm according to the group-1 rather than the group-2 sequence? Cross-tab analyses answer this question, as they identify whether entrepreneurs that belong to one group also vary systematically in other characteristics, such as their gender, their motive to start a company, or their age. One could, for example, imagine systematic co-variations between founding motives and founding sequences, where necessity entrepreneurs are more likely to open a company according to the group-1 sequence, whereas opportunity entrepreneurs proceed according to the group-2 sequence.

Criticism: The Decisions to be Taken for Conducting SA

Amongst natural scientists, the usefulness of SA is uncontested ever since this method was employed to analyze DNA sequences and decode the human genome (Abbott and Tsay 2000). Social scientists are however less convinced about the added value of SA. While some acknowledge that this method can be a useful tool for data assessment (see MacIndoe and Abbott 2004; Abbott and Hrycak 1990), many social scientists remain sceptical that SA provide meaningful insights (see, for example, Levine 2000; Wu 2000). The criticism against SA can be summarized along the lines of the steps in which SA are carried out.

With regard to the **first step** (ordering of sequences), critics point to the difficulty of taxonomy (Levine 2000; see also Elzinga 2003: 5). More concretely, this difficulty comes in two forms: first, researchers need to decide whether, or not, to report the duration and, hence, the timing of events. In other words, shall events or only states be reported? Second, and more problematically, the question arises how to deal with incomplete data. As with most datasets, also SA data usually contains incomplete information, i.e. sequences that are unfinished or contain gaps. In these instances, the researcher does not know whether an event took place and, if so, which one. Take the sequence `V-I-?-?-E`: what happened after the entrepreneur had gathered information and before he hired employees? Did he undertake one or even two founding steps about which the researcher has failed to ask? Or did he discontinue all founding activities for two periods? Akin to traditional statisticians, researchers intending to conduct SA need to decide how to treat (sequences with) incomplete data. In essence, they have two options. SA users can either cut out the unknown states, which implies that they shorten the overall sequence. They can, for example, cut out all unknown events of the `V-I-?-?-E` sequence and shorten it to `V-I-E`. Alternatively, researchers can decide to leave incomplete sequences intact and treat unknown states as if the missing information `?` was one additional event. Hence, supposed `O` stands for `other activity`, the sequence `V-I-?-?-E` could be transformed into `V-I-O-O-E`. Note, however, that Abbott and Tsay (2000) illustrate how misleading results can be obtained if unknown events are treated like known events and assigned the same transformation costs.

With regard to the **second step** (calculation of distances), critics highlight two difficulties. First, it is often unclear which algorithm is the most useful one to analyze a given dataset (see Dijkstra and Taris 1995; Elzinga 2003; Prinzie and Van den Poel 2006; Lesnard 2006). This is particularly true as some algorithms allow only for the substitution but not for the insertion and deletion of individual events. Hence, the question arises which algorithm to choose. Second, possibly the most prominent difficulty of conducting SA is the specification of transformation (i.e. substitution and indel) costs. For successful SA analyses of social science data, it is both particularly important and challenging to set transformation costs (see Abbott and Tsay 2000: 12). On the one hand, the sheer number of transformations for which costs need to be specified can be huge (remember that the substitutions of A by B and B by A are two different processes). On the other hand, it is necessary to set asymmetrical costs whenever the transformation of one process into another is less likely than the opposite transformation. It might, for example, be more difficult to acquire venture capital (V) than hire employees (E) so that the substitution of `E` by `V` is more costly than the substitution of `V` by `E`. Should a researcher lack the necessary knowledge of cases to set asymmetrical transformation costs, SA is likely to produce less meaningful results (Wu 2000; Prinzie and Van den Poel 2006).

In line with step 2, researchers also find it difficult to identify the most meaningful algorithm for conducting cluster analyses in **step 3** (clustering) (see, for example, Jain, Murty et al. 1999; Zhao and Karypis 2002; Prinzie and Van den Poel 2006).

In view of the numerous choices that social scientists need to make in order to conduct SA, it is hardly surprising that many find this method confusing. While most social scientists do not reject SA *per se*, they rather have difficulties to identify the most meaningful way of running it. How to proceed in order to understand whether SA yield insightful results? The remainder of this paper seeks to answer this question. To this end, we develop a decision tree that recapitulates all those questions that researchers need to address in order to carry out SA. In essence, four choices need to be made. To complete step 1, researchers need to decide whether, or not, to report the length of individual events and, consequently, entire sequences (report length of sequences?). Furthermore, and still related to step 1, they need to decide how to treat incomplete data sequences (what to do with incomplete data?). With regard to step 2, researchers must choose an SA algorithm and set transformation costs (which algorithm to choose and how to set transformation costs)? Finally, to complete step 3, they need to choose an algorithm for conducting cluster analyses (which algorithm for cluster analyses to choose)? While our decision tree allows researchers to make their options explicit, we do not propose to choose just one way of running SA. Instead, we recommend to always `do it both ways` (section 3.3). Depending on the data available and the knowledge of this data, researchers can experiment along the lines of the decision tree in order to understand whether and, if so, which SA processes deliver meaningful answers to their research questions. To illustrate how our decision tree can be used, let us turn to a real-world dataset on venture creation processes.

3. `HYPOTHESES`: THE SA DECISION TREE OR `DO IT BOTH WAYS`!

For three reasons, the study of venture creation processes constitutes a particularly insightful case to illustrate the usefulness of SA in general and the use of our decision tree in particular. First, agreement amongst social scientists is broad that entrepreneurship is a driver of economic growth (Audretsch and Thurik 2001; Carree and Thurik 2003; Stel, Carree et al. 2005). Consequently management and business researchers, sociologists, psychologists, economists, innovation scholars, and political scientists share an interest in understanding how entrepreneurship evolves. Second, agreement among entrepreneurship scholars is broad that this phenomenon ought to be studied from a dynamic perspective. As Gartner (1988) skillfully argues in his path-breaking article: even if all traits (such as gender, age, educational background etc.) of the ideal-typical entrepreneur were known, a person with all these traits could very well not be an entrepreneur. This leads Gartner and his followers to argue that studies of venture creation processes teach us more about how entrepreneurship materializes than studies of entrepreneurial traits. Third, and in striking contrast to the two aforementioned agreements, entrepreneurship scholars have to date not conducted SA of venture creation processes – with two, fairly one-sided exceptions. Analyzing financial trajectories of organizations with the help of SA, Keister (2004) identifies four paths that Chinese firms take to raise funds. In a similar vein, Garnsey et al. (2006) investigate the growth path of firms in the Netherlands, Germany, and the UK. Distinguishing between three phases of organizational development (growth, stagnation, and decline), the authors use SA and identify four archetypical development paths. Since the authors of both studies conduct SA to complement the insights gained from traditional statistical methods, they ultimately tell us little about the use and

usefulness of SA for analyzing social science data in general and venture creation processes in particular.

Aiming to fill this gap in the entrepreneurship and SA literature alike, we carry out SA analyses of the largest available dataset on venture creation processes: the Panel Study of Entrepreneurial Dynamics (PSED). In doing so, we follow up on the criticism raised against SA and make the choices of SA analysts, as well as their respective implications, explicit. This ultimately leads us to propose a decision tree which recapitulates 16 variants of how to run SA. The results presented in section 4 tell the reader how to identify the most opportune out of all 16 variants.

3.1. The Data: The Second Panel Study of Entrepreneurial Dynamics

To date, the most comprehensive and most recent database on venture creation processes is the second Panel Study of Entrepreneurial Dynamics (PSED2),[2] which was compiled on an annual basis between September 2005 and April 2009. Coordinated by scientists at the University of Michigan, more than 100 researchers at 36 US universities contributed to data collection. Overall 1,214 entrepreneurs were selected from a pool of more than 31,000 adults and interviewed by mail or phone in four repetitive survey waves. Information was collected about 38 start-up activities and several hundred entrepreneurial attributes including, *inter alia*, the composition of the start-up team, reasons for opening the business, its technological intensity, industry and geographical location.[3]

Even though the PSED2 data is the most comprehensive database on venture creation processes, it traces starting and completion dates for only 5 out of the overall 38 founding activities observed. These five events are: acquiring intellectual property (IP), acquiring outside funding (OF), development of service or product (DP), investments made by owners (IO), and writing a business plan (BP). For the remaining 33 activities only the completion date was measured, allowing for comparisons of the type and order, but not the length of activities. Importantly, and unfortunately for dynamic entrepreneurship research, this disqualifies even the PSED2 data as an empirical basis for identifying complete founding sequences. Nevertheless, the subset of observations including those 5 founding steps, for which start and end dates are available, is large enough to illustrate how SA are carried out in general and how these analyses can provide meaningful insights into venture creation processes. One might even argue that the imperfections of the PSED2 data make it an ideal-typical case for exploring the usefulness of SA, because datasets are hardly ever complete or otherwise `perfect` for the use of any given method. Table 1 gives an overview of all venture creation activities included in the PSED2 dataset and also reports their level of measurement.

---

[2] We here use the second rather than the first PSED database – i.e. PSED2 rather than PSED1 – because the latter is both less comprehensive and less up-to-date than its successor. For similar reasons, we also decided not to use the KEINS (Knowledge-Intensive Entrepreneurship Innovation, Networks and Systems) database: the Europe-based counterpart to the US-based PSED studies (see Lissoni, Sanditov et al. 2006). Contrary to the PSED, the KEINS data does not provide start and end information of individual venture creation activities so that start-up sequences cannot be studied from a dynamic perspective.

[3] For a detailed description of the data and interview protocols see Curtin (2009).

Table 1: Venture creation activities traced in the PSED2 dataset (alphabetical order)

Gestation activity		Gestation activity	
Accountant retained	*completion date only*	Financial projections	*completion date only*
Acquiring intellectual property (IP)	***start and end date***	First personnel hired	*completion date only*
Acquiring of mayor equipment	*completion date only*	Found additional financial support	*completion date only*
Acquiring outside funding (OF)	***start and end date***	**Investments made by owners (IO)**	***start and end date***
Acquiring raw materials	*completion date only*	Lawyer retained	*completion date only*
Application for 'doing business as'	*completion date only*	Legal form registered	*completion date only*
Application for federal EIN number	*completion date only*	Liability insurance purchased	*completion date only*
Bank account first used	*completion date only*	Listed with Dunn and Bradstreet	*completion date only*
Began working fulltime	*completion date only*	Marketing efforts started	*completion date only*
Business name registered	*completion date only*	Membership of Trade Association	*completion date only*
Collecting competitor information	*completion date only*	Partner becomes involved	*completion date only*
Customer discussions	*completion date only*	Physical space first used	*completion date only*
Defined market opportunities	*completion date only*	Registered in phone book or internet	*completion date only*
Determined regulatory requirements	*completion date only*	Revenues exceeded expenses	*completion date only*
Developed proprietary technology	*completion date only*	Signed formal agreement ownership	*completion date only*
Developed service or product (DP)	***start and end date***	State unemployment Insurance	*completion date only*
Entrepr.eur ended active role in firm	*completion date only*	Supplier credit established	*completion date only*
Federal income tax return filled	*completion date only*	Thinking of new business	*completion date only*
Federal social security payed	*completion date only*	**Writing business plan (BP)**	***start and end date***

Source: Overview based on PSED2 questionnaires
(available at http://www.psed.isr.umich.edu/psed/data)

3.2. SA Analyses: The Decisions to be Made

Using the PSED2 data as an empirical basis, let us go through the individual steps of an SA analysis and recapitulate, for each step, which decisions need to be taken.

Step 1: Building Sequences

As laid out in section 2, the first two decisions to be taken concern the question of how to build meaningful sequences. Answering this question implies to decide first whether, or not, to consider the *duration* of individual events. In other words, shall sequences be reported as `State Sequences` (STS) or as `Distinct State Sequences` (DSS)? While STS report the type and order of events (states) as well as their duration, DSS allow events to have a duration of one time period only. Hence, DSS describe the type and order of events without considering their duration. Table 2 summarizes the difference between the two ways of building sequences.

Table 2: Differences between STS and DSS sequences

Sequence Type	Type of Data Required	Information Reported	Example
STS (State Sequences)	Start and End Date	Type, order, duration	OF-?-?-?-OF-OF-DP-IP-IP -IP
DSS (Distinct State Sequences)	Start or End Date	Type, order	OF -?- OF - DP -IP

With regard to our analyses of the PSED2 data, the most important implication of table 2 is the following. Whenever sequences are reported in the DSS format, it is sufficient to have only a starting or a completion date of an event, whereas events can only be reported in an STS format if both their starting and completion date is known. In other words, we can build sequences in the DSS format including all 38 events traced in the PSED2 database, whereas we can build STS sequences only on the basis of those 5 events for which both the start and end date is known. It should be noted that we decided to build STS *as well as* DSS sequences only on the basis of the 5 complete events. The reason for not including the other 33 events in DSS sequences is straight-forward. We ran SA for DSS sequences with both 38 and 5 events, and we found that the results do not differ substantially. In order to facilitate the comparability of the SA results obtained for DSS and STS sequences, we decided to build not only STS but also DSS sequences out of the 5 events for which start and end dates are reported. Overall, data on at least 1 of these 5 activities was available for 1125 venture creation sequences of the overall 1214 sequences included in the PSED2 database. Hence, we started our SA analyses of the PSED2 data with building 1125 venture creation sequences in the STS as well as in the DSS format.

The second choice to be made when building sequences concerns the question what to do with incomplete data. To begin with, we decided to take the first founding activity undertaken by an entrepreneur as the starting date of the venture creation process. Similarly, we considered a founding process as completed once the final founding activity was ended. In this way, all sequences have a clear beginning and end date. We also decided to use the same time interval as the PSED2 and report all sequences in intervals of one month. Regarding gaps within the sequences – for which it is unclear whether and, if so, which founding steps were undertaken – the decision to be made is whether, or not, to report event gaps. Akin to the choice between STS versus DSS formats, we decided to do it both ways. Hence, we built 1125 sequences including the event gaps (blanks), and 1125 sequences excluding the blanks.[4] Note that the latter process does not only shorten the overall length of sequences, but also changes the relative timing of events as they occur earlier than in sequences which retain event gaps.[5]

[4] For a better understanding of the process that led to propose a decision tree, note that we actually built 1125 STS sequences including blanks and 1125 STS sequences excluding blanks, *as well as* 1125 DSS sequences including blanks and 1125 DSS sequences excluding blanks.

[5] Note that we explore only single-channel, but no multi-channel SA in this paper. In other words, we only use algorithms that can process one event at a time. This implies that we needed to modify those, rather few, sequences in which two or more founding events took place at the same time. In these cases we used a `last come, first served` approach. In order to maintain the overall length of a sequence, we worked backwards from the last event, which we maintained as the end date of the founding process. For any earlier time period, we then considered whether two events took place at the same time. Whenever this was the case, the event that ended later would overwrite the other

Step 2: Distance Calculation

The next choice to be made concerns the algorithm for distance calculation. Which algorithm to choose for calculating distances between sequences? In essence, this question comes down to deciding whether sequences shall be compared according to the *types* or, rather, the *order* of their events. Remember that some algorithms only use substitutions while others use indel operations in addition. While the former give priority to preserving the order, the latter give priority to preserving the types of events.

From the existing SA algorithms, two have proven to be particularly powerful for the majority of sequence data: the Optimal Matching (OM) and the Longest Common Subsequence (LCS) algorithms. Most fundamentally, they differ in that LCS only uses indel operations, whereas the OM algorithm employs substitution as well as indels. Depending on how transformation (substitution and indel) costs are set, the OM algorithm thus makes it possible to attach equal importance to comparing the *types* and *order* of sequence events. Due to this flexibility, OM has become the standard algorithm for SA analyses of social-science data (Abbott and Tsay 2000). Yet, the LCS algorithm has proven to be a particularly powerful alternative. Since it only makes use of indel operations, prioritizing the order over the types of events, the LCS algorithm produces particularly insightful results whenever the length of individual sequences differs considerably (Elzinga 2003). Given that start-up processes were found to vary greatly in length (between 1 month and 10 years) (Reynolds and Miller 1992; see also Reynolds and Curtin 2008: 263), the LCS algorithm seems the most useful alternative for studying venture creation processes.[6]

---

founding event(s) that took place in parallel, but ended earlier. For example, a sequence with the two parallel founding activities:

1) * - * - * -OF-OF-DP-IP-IP-IP and

2) BP-BP-BP-BP

would be combined into one single sequence as follows:

1+2) BP-BP-BP-OF-OF-DP-IP-IP-IP.

Whenever two events ended at the same time, the parallel states were overwritten in the following order: BP overwrites OF overwrites IP overwrites IO overwrites DP. Consequently, a sequence with the two parallel founding activities:

1) * - * - * -OF-OF-DP-IP-IP-IP and

3) BP-BP-BP-BP-BP

would be combined into one single sequence as follows:

1+3) BP-BP-BP-BP-BP- DP-IP-IP –IP.

[6] Depending on the type of data to be analyzed, researchers might find several other algorithms particularly useful tools. They include: (1) *The Longest Common Prefix* (LCP) and (2) *Longest Common Suffix* (RLCP) algorithms (Gabadinho, Ritschard et al. 2008). These algorithms are particularly suitable whenever the event patterns of interest are located at the beginning or the end of the respective sequences. (3) The *Hamming Distance* (HAM) algorithm uses only substitutions, which makes it an opportune method for analyzing sequences of equal length (Elzinga 2007). (4) The *Dynamic Hamming Distance* (DHD) is similar to the HAM algorithm but makes it possible to set variable substitution cost at different positions of the algorithm (Lesnard 2006). (5) The *Asymmetrical OM* algorithm is identical to the standard OM in that both use the weighted Levenshtein algorithm to calculate distances between sequences. However, asymmetrical OM make it possible to set transformation costs with the help of an asymmetrical substitution matrix, so

In addition to choosing algorithms, SA researchers also need to decide how to set transformation costs. In section 2, we highlighted the importance of setting substitution and indel costs that represent the `symmetry` of event transformations. If it is more difficult for business owners to acquire outside funding (OF) than to invest their own funds (IO), then the substitution of OF by IO should be less costly than the substitution of IO by OF. So much for the theory (for an exhaustive overview of how to set transformation costs, see Abbott and Hrycak 1990; Abbott and Tsay 2000). In practice, however, researchers do typically not know about asymmetries of transformations whenever they analyse datasets which they have not compiled themselves. Our use of the PSED2 data is a good example. Since we lack the necessary information to set transformation costs either manually or via a fixed hierarchy, we assign costs automatically with the help of a transition matrix (see MacIndoe and Abbott 2004).

In sum, we decided to analyse our STS and DSS sequences, including and excluding blanks, with the OM and LCS algorithm, whereby transformation costs were always set automatically.

3. Step 3: Cluster Formation

Once distances between sequences have been calculated, the final question to be addressed is how to cluster sequences. How to identify the most *meaningful* number of clusters: i.e. how to identify *that* number of clusters, for which sequence differences *within* one cluster are minimized, whereas sequence differences *between* clusters are maximized? Note that this, most meaningful, number of clusters is identified in two steps. While different sequence clusters are built (cluster formation) in a first step, their quality is assessed in a second step (cluster validation). Since the cluster validation process (step 2) also gives an answer to the overarching question of whether SA provide meaningful insights into social-science processes, we discuss cluster validation in the results section (section 4).

In this section, we focus on discussing the decisions to be made for cluster formation. How to divide the entity of sequences (in our case 1125 venture creation processes) into clusters? To answer this question, a researcher needs to choose a clustering algorithm, whereby a fundamental difference exists between *hierarchical* and *partitional* clustering methods[7]. *Hierarchical clustering algorithms* are based on the assumption that social processes evolve, like an `evolutionary tree`, from one primary sequence (Kaufman and Rousseeuw 1990). Consequently, the hierarchical clustering algorithm starts from one end of the data spectrum. It either divides an all-encompassing cluster into more appropriate ones (divisive hierarchical clustering), or it merges all individual data points (sequences) stepwise into more encompassing clusters (agglomerative hierarchical clustering). The most powerful hierarchical clustering algorithm – predominantly used in SA research to date (see Abbott and Hrycak 1990; MacIndoe and Abbott 2004; Brzinsky-Fay 2007; Muller, Gabadinho et al. 2008) – is named after its founding father Joe H. *Ward* (Ward Jr 1963).

---

that unequal transformations between two events can be described (Prinzie and Van den Poel 2006). Of course, any of these algorithms can be used as an alternative for the OM or LCS algorithm we used, or as the basis for developing a new branch of the SA decision tree (see section 3.3).

[7] More recently, *self organizing maps* were used as a further clustering algorithm for SA (Massoni, Olteanu et al. 2009). But since this method is still in its infancy, only hierarchical and partitional clustering algorithms are discussed here.

Ward`s clustering algorithm has not only been used with notable success in biology research, it is also appreciated for its comparatively short computation time.

Partitional clustering algorithms were developed more recently, in response to the criticism that processes do often *not* evolve from one primary sequence. Consequently, partitional clustering does not start from either one all-encompassing or atomized set of clusters. Instead, the researcher predefines a number of clusters from the outset, into which the entire dataset is *randomly* divided in a first step. In each subsequent step, data points (sequences) are added or removed from these clusters to improve their overall fit. Due to the random starting situation, partitional clustering methods can deliver slightly different results each time the algorithm is run. Moreover, computing time is comparatively high. However, partitional clustering methods have the advantage that they offer powerful tools to visualize SA results (see Massoni, Olteanu et al. 2009). Furthermore, and more importantly, partitional clustering algorithms have proven to yield similar or even better results than hierarchical clustering methods. This is particularly true for large datasets including highly diverse cluster (Zhao and Karypis 2002; Zhao and Karypis 2005). Given that venture creation sequences have been found to vary greatly in the duration and order of individual founding activities (Reynolds and Miller 1992), partitional clustering algorithms seem particularly useful for SA analyses of the PSED2 dataset. The most powerful algorithm of this family is the *Partitioning around Medoids* (PAM).

In line with our previous choices, we again decided to do it both ways and to cluster our STS and DSS sequences – including and excluding blanks, using OM and LCS algorithms for distance calculation – with the Ward and the PAM algorithm.

3.3. The SA Decision Tree: `Always Do It Both Ways`!

When recapitulating the individual steps of SA and the choices related to these steps, the author had to think back of her times as a PhD student at the European University Institute. At that time, one of the economics professors had pinned a widely visible postulate at his office door: `Do it both ways!`. His supervisees credibly assured us that they received this piece of advice with beautiful regularity. Our experience with SA leads us to give a similar advice, as we think that SA researcher should `always do it both ways`.

Even if researchers analyze a dataset which they collected themselves, it is often hard to judge from the outset, whether the more relevant information is preserved: if sequences are reported in a STS or DSS format, with or without events gaps, if sequence distances are calculated with the OM or LCS algorithm, and if clusters are formed with the Ward or the PAM algorithm. Based on this insight, we developed the decision tree depicted in figure 1.

It should be clear from our previous illustrations that this decision tree is not set in stone. While the SA steps and the *types* of choices to be made are always the same, the *number* of choices can vary notably, depending on the dataset to be studied. The composition of the PSED2 data indicated that 16 different ways of running SA may deliver particularly promising results. Other datasets could however make it more useful to set substitution costs not only automatically but also manually. This would increase the number of potentially insightful SA to 32. Similarly, researchers may decide to explore only one, or even three, rather than two cluster algorithms. This would cut one branch from, or add a new branch to, the decision tree respectively. The central message behind the decision tree therefore is: in case of doubt, do it both ways!

Figure 1: The SA Decision Tree

Step 1:

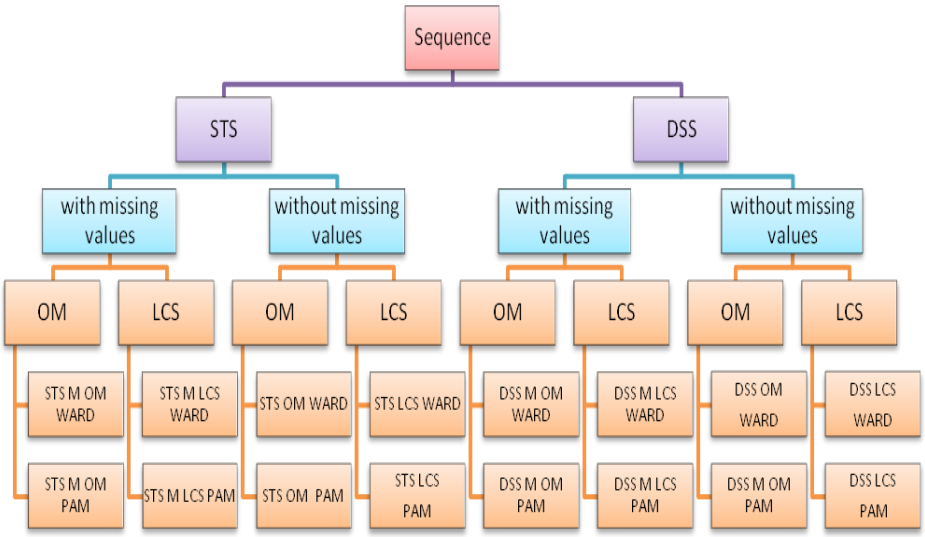
1. Consider length of individual sequence events?
2. What to do with incomplete data?

Step 2: Distance Calculation

3. How to set transformation costs ➔ which algorithm to choose?

Step 3: Clustering

4. Which cluster formation method to choose?



4. RESULTS

The results that can be obtained from applying our decision tree are of two kinds. First, the ideal-typical ways of setting up a company, including their numbers and most common sequences, can be identified. This is what we will do in section 4.1. Furthermore, the founding sequences identified can be explained with the help of cross-tab analyses: our task of section 4.2.

4.1. Cluster Validation: How to Identify the `Best Way`

As has been pointed out at the end of section 3.2, sequence clustering usually involves one further step, namely the validation of clusters. The reason is that both hierarchical and partitional algorithms can divide a dataset into any given number of clusters. However, they do not identify the most meaningful cluster number, i.e. that number for which sequence differences within clusters are minimized while they are maximized between clusters. Hence, whenever researchers cannot, or do not want to, predefine a number of clusters *a priori*, they need to identify the most meaningful number of clusters *ex post* with the help of cluster validation. Our analyses of the PSED2 data provide a good example. Since we are not interested in identifying venture creation sequences for a given number of clusters, but since we rather want to understand how many ideal-typical founding processes can be identified, we need to assess the most meaningful number of clusters *ex post* through cluster validation.

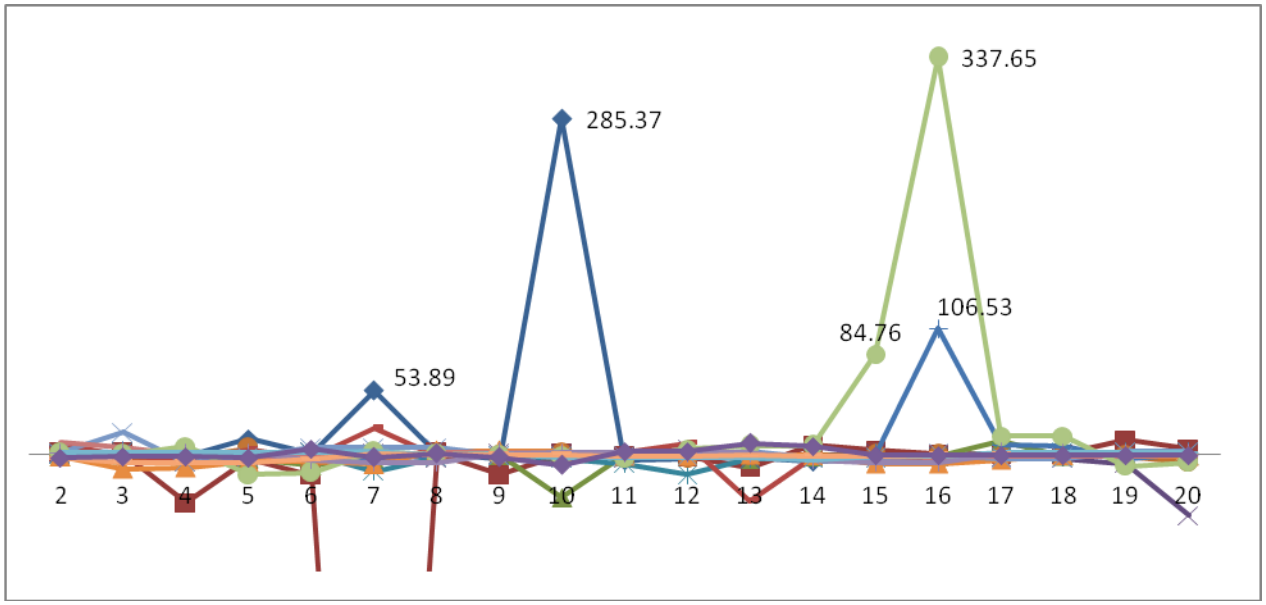
It might not come as a surprise to the reader that several cluster validation algorithms exist. In their highly influential article, which is still considered the standard reference today (see Savova, Therneau et al. 2006), Milligan and Cooper (1985) compare 30 different cluster validation procedures and evaluate their performance. For the sake of simplicity, suffice it to say here that we chose that cluster validation algorithm which Milligan and Cooper consider the most appropriate one for our type of data and research aim: the gamma index.

How does cluster validation work in general, and what does the gamma index teach us in particular? To begin with, cluster validation algorithms calculate a `goodness-of-fit` score for any given number of clusters. Take our example of venture creation processes, for which we decided to group the overall 1125 PSED2 sequences into 2, then 3, then 4 (etc.), and finally 20 clusters. Note that this was still done with the help of the Ward and PAM *formation* algorithms. For each number of clusters (in our case, 2 to 20), *validation* algorithms then calculate a score which indicates the `goodness of fit`. The latter depends on how the distance between clusters is calculated. Yet, any validation index has an upper and a lower bound of scores that it can assume. The gamma index calculates cluster distances by dividing the number of consistent outcomes (i.e. those observations that are close and within the same cluster) by the number of inconsistent outcomes (i.e. close but in other clusters). Consequently, the index`s lower bound is $-\infty$, while the upper bound is $+\infty$. The closer a goodness-of-fit score is to $-\infty$, the worse is the cluster quality, because many data points (sequences) that are similar to each other are not situated in the same cluster due to an unrepresentative number of clusters chosen. Vice-versa, high gamma scores indicate that the number of groups chosen divides data points (sequences) into meaningful clusters, where similar sequences are located in the same cluster. Consequently, the number of groups for which the highest gamma score is obtained divides sequences into the most meaningful clusters.

Let us apply this approach to identify the most meaningful SA path. Remember that we have thus far discerned 16 potentially useful ways of running SA and, for each of these ways, we

decided to form 2 to 20 clusters (i.e. 19 groups). For these overall 304 (= 16 * 19) clusters, we now calculate a goodness-of-fit score with the help of the gamma index. The interpretation of the results obtained is straight-forward: that SA path with the highest score indicates the most meaningful cluster division and, hence, the most meaningful way of running SA. Figure 2 graphs the outcome of the goodness-of-fit scores for all 16 ways of running SA and the 19 different clusters into which the PSED2 dataset was divided.

Figure 2: The most meaningful way(s) of running SA (goodness-of-fit scores of gamma index)



Colour charts:

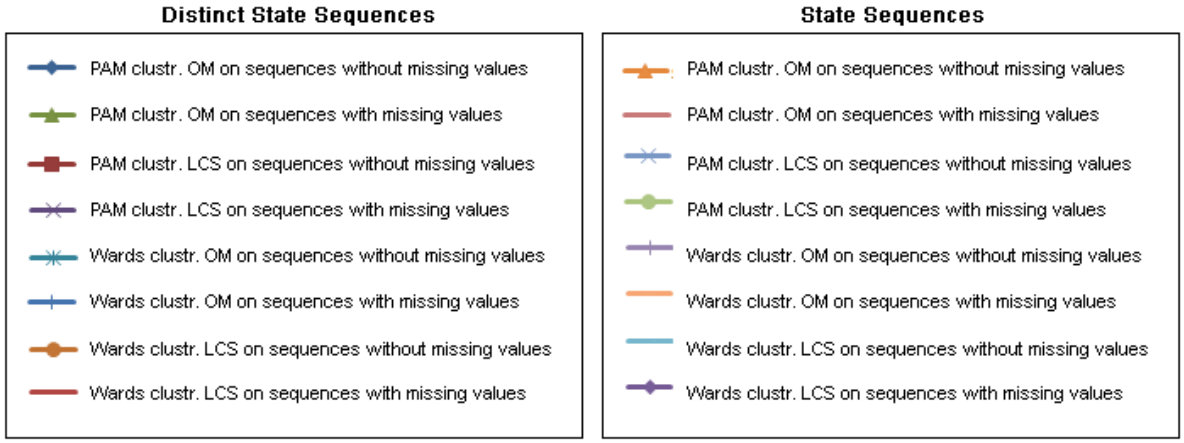
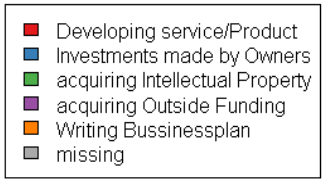


Figure 2 clearly illustrates that different ways of running SA produce highly diverse outcomes. Most importantly, one gamma index – resulting from SA analyses of STS sequences including

event gaps, compared to each other with the LCS algorithm and clustered into 16 groups with the help of the PAM algorithm – is so high (337.65 points) that it outperforms all other scores. In essence, this result tells us that there are 16 different and ideal-typical ways of creating a new venture. Note that it is neither necessary nor usual that one way of running SA turns out to be so clearly superior to all other ways of identifying common sequence patterns. Supposed that venture creation was a purely random process, none of the clusters would have looked particularly different from their counterparts – irrespective of the SA procedure used.[8] In sum, the decision tree has delivered a clear-cut outcome. SA analyses that use the LCS algorithm for distance calculation in order to cluster STS sequences including event gaps into 16 groups with the PAM algorithm deliver most meaningful clusters.

It should be clear from our previous illustrations that the most representative venture creation sequence can be identified for any of these 16 clusters. And, of course, we want to know how these sequences look like. They are easiest to discern when they are visualized. While a *frequency plot* visualizes all sequences included in one cluster, a *state distribution plot* reports the sequences of event frequencies for each time period, whereas a *representative sequence plot* graphs the most common sequence of the cluster. For the sake of space, we shall here only reproduce some of the most insightful sequences. They include `the short-lived enterprises` of cluster 1 (comprising 608 entrepreneurs) for which the most common founding sequence is one month of investments made by the owner (figure 4). Furthermore, cluster 3 delivered particularly insightful results as it includes 39 `business plan writers` who focus on writing a business plan for three months, then try to acquire outside funding for two months, and finally continue writing business plans for another nine months (figure 5). Also the 51 `product developers` of cluster 4 pursue a particularly unique founding path as they invest their own funds for one month in order to develop a product or service for the following six months. After that, they again invest own funds for one month so as to continue product or service development for another four months (figure 6). Finally, cluster 7 constitutes an interesting contrast to cluster 1 as it includes 27 `investors` whose most-common activity is the investment for their own funds for an overall period of 15 months (figure 7).

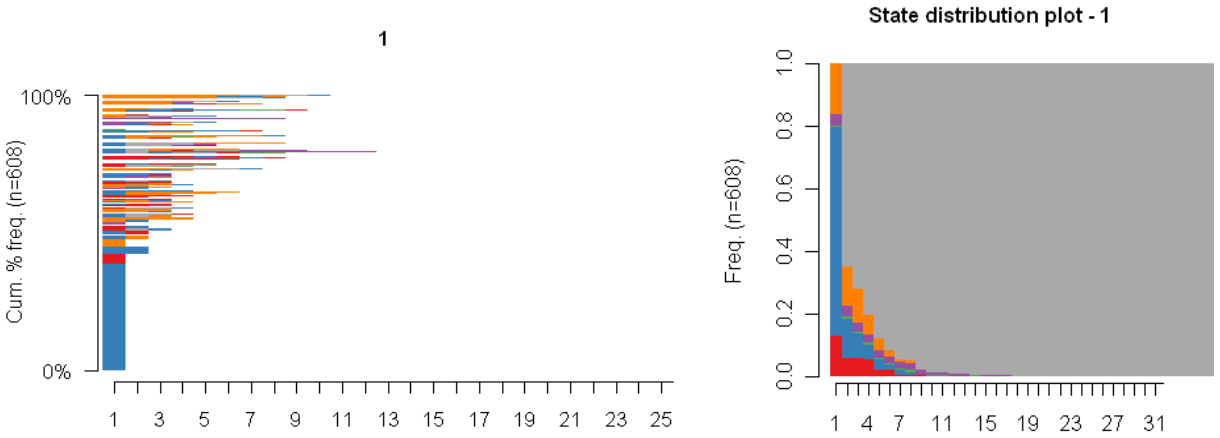
Figure 3: Colour Chart Used in Figures 4-7

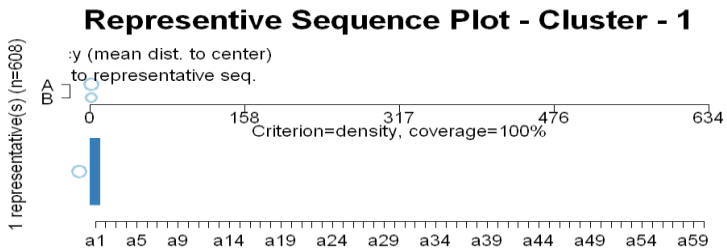


[8] In fact, most methods yield fairly low scores: 132 out of 304 clusters score between -10 and 10 on the gamma index. Out of those ten clusters with a gamma score above 10, four clusters were produced by SA analyses of STS sequences including event gaps for which differences were calculated with the LCS algorithm and clustered into groups using the PAM method. This approach clearly outperforms all other SA types as it produces a gamma score of 84.76 for 15 clusters and the aforementioned 337.65 for 16 clusters. Only SA analyses of DSS sequences excluding event gaps, compared with the OM algorithm and clustered with the PAM method yield gamma indices that come close to this score: of 285.37 for 10 clusters and 53.98 for 7 clusters. None of the other methods score higher than 21.98 for any given cluster size.

Cluster 1: The Short-Lived Enterprises

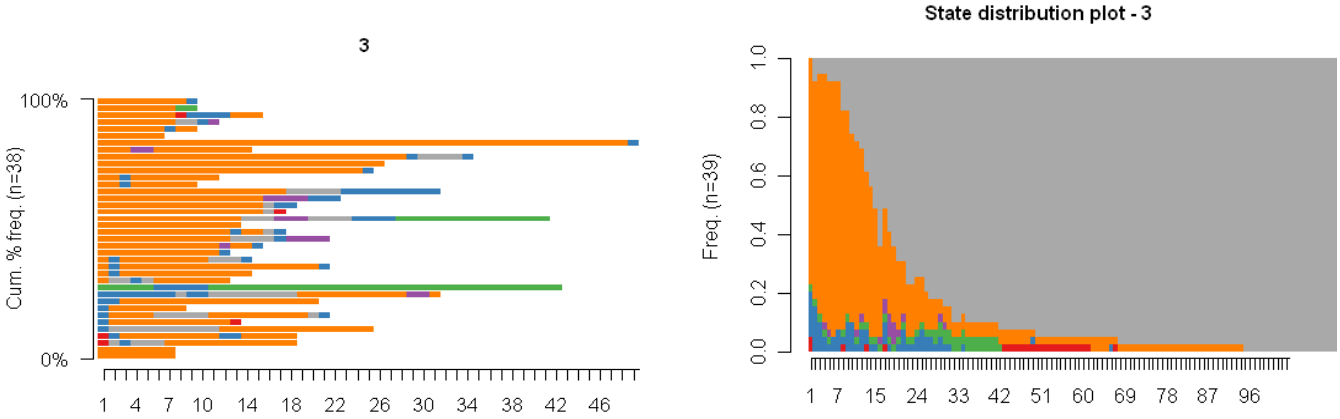
Figure 4: *Frequency plot, state distribution plot,* and *representative sequence plot* of cluster 1

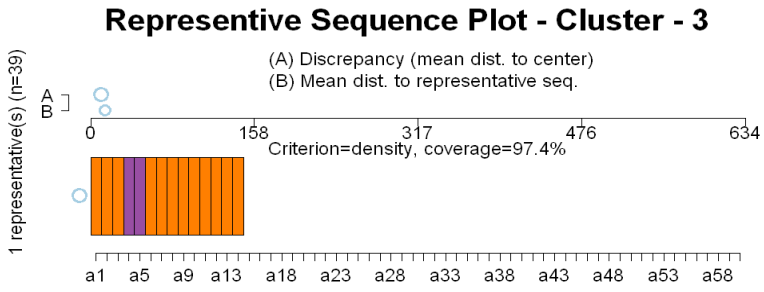




Cluster 3: The Business Plan Writers

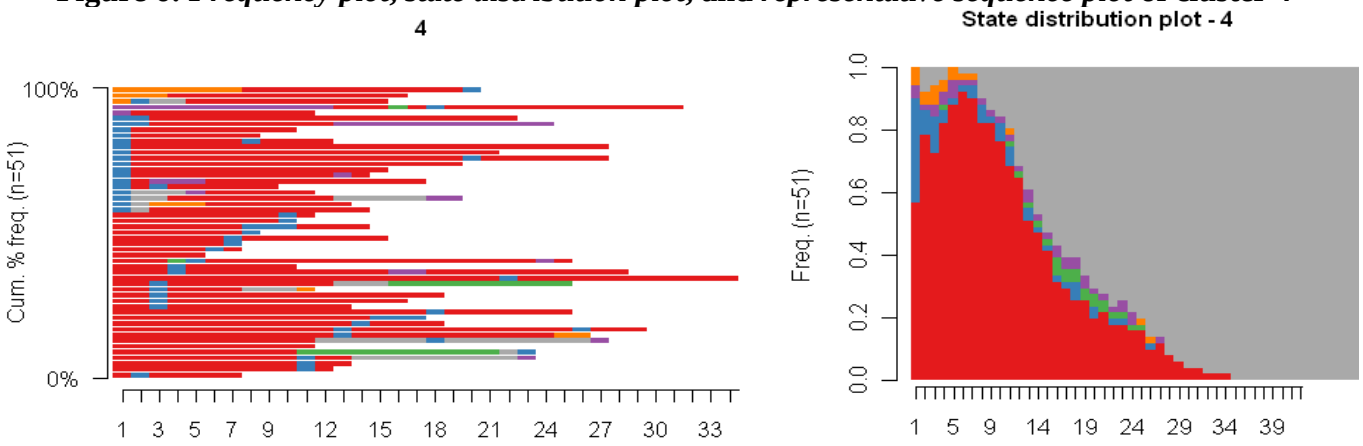
Figure 5: *Frequency plot, state distribution plot,* and *representative sequence plot* of cluster 3

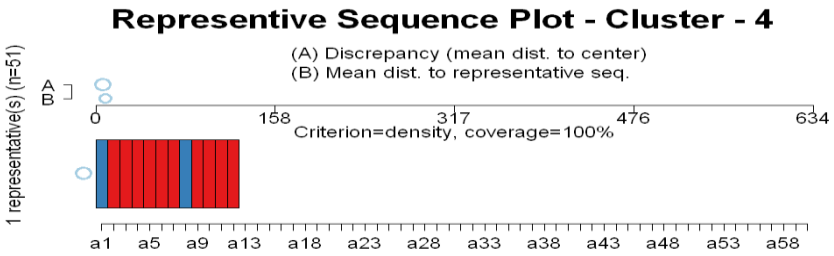




Cluster4: The Product Developers

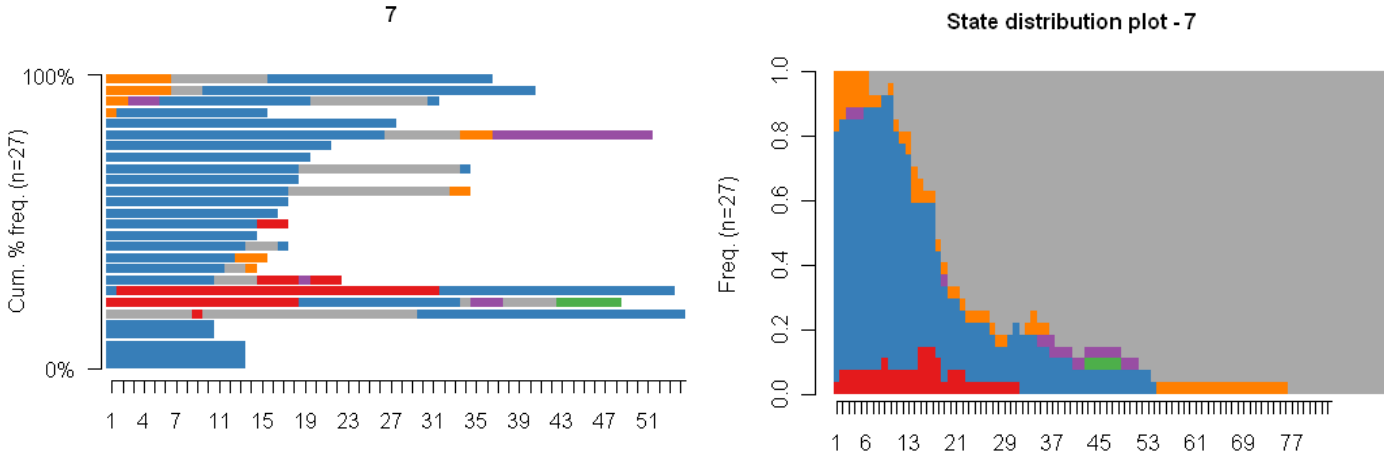
Figure 6: *Frequency plot, state distribution plot,* and *representative sequence plot* of cluster 4

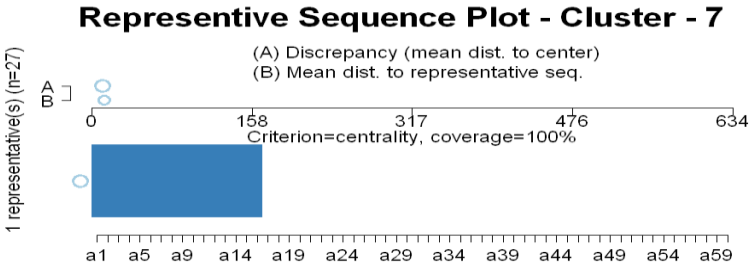




Cluster 7: The Investors

Figure 7: *Frequency plot, state distribution plot,* and *representative sequence plot* of cluster 7





4.2. Understanding SA Results: How To Explain Venture Creation Processes

While the16 founding sequences identified with SA are insightful per se, as they tell us how many ideal-typical founding sequences exist and how these look like, we still do not know *why* sequences differ. In other words, which factors lead entrepreneurs to set up their company in one rather than another way? To answer this question, note how we highlighted in section 2 that SA is often used in combination with other statistical tools, most frequently with cross-tab analyses. Accordingly, we argued that the fourth and final step of SA consists in explaining differences between clusters with the help of cross tabs.

For venture creation processes, the following factors seem particularly likely to influence the venture creation path of an entrepreneur (Gartner, Shaver et al. 2004): the founding *team* (solo entrepreneur versus group of people), the *start-up motive* (necessity versus opportunity situation), the *industry* of the new venture (measured in SIC codes at the 1-digit level), the founding *region* of the venture (measured in compass and census divisions), the venture`s *origins* (spin-off versus independent start-up), its *location* (urban, suburban or countryside), its *legal form* (sole proprietorship, general partnership, limited partnership, limited liability corporation or general corporation), its *technological intensity* (high- versus low-tech), its *profitability* (profit versus no profit made), and the *revenue* generated (higher or lower than expenses). Data for all these variables were available from the PSED2 database for up to 1125 cases.

Table 3: Explanations of venture creation processes (cross-tab analyses)

Variable	Pearson`s Chi^2	Cramer's V	Low cell count [2)]	No. of clusters [3)]	No. of cases (sequences)
Team	56.032***	0.158***	0.0 %	11	1125
Start-Up Motive	21.149**	0.148**	0.0 %	11	971
Industry	40.056**	0.094**	8.6 %	5	1125
Founding Region					
9 census divisions	66.774**	0.099**	17.5 %	7	1083
4 wind regions	36.926	0.105	4.5 %	11	1125
Origin (Spin-off)	16.772*	0.122*	9.1 %	11	1122
Location (Urban)	42.822	0.098	18.2 %	11	1125
Legal form	18.047	0.068	12.0 %	5	964
Technol. Intensity	6.491	0.076	0.0 %	11	1124
Profitability	3.616	0.107	0.0 %	5	314
Revenue	14.579	0.114	0.0 %	11	1125

Notes:
1) Significance levels: * < 0.10; ** < 0.05; *** < 0.01.
2) Low cell count: A maximum of 20% of cells with an expected cell count < 5 is considered acceptable.
3) Clusters in analysis:
- 11 cluster solution = cluster 1, 2, 3, 4, 5, 6, 7, 9, 10, 12 and residual (see footnote 9)
- 7 cluster solution = cluster 1 , 2, 4, 5, 9, 12 and residual (see footnote 9)
- 5 cluster solution = cluster 1, 2, 5, 9 and residual (see footnote 9)

Table 3 reports the results obtained from cross tabulations of the sequences we identified[9] and the aforementioned variables. Let us begin with discussing those factors that do *not* influence the ways in which entrepreneurs proceed to start a business. As indicated by the statistical insignificance of the Chi-Square and Cramer`s-V scores, venture creation processes are neither influenced by the region (North, East, South, West) in which an enterprise is started, nor by the venture`s location with regard to large cities. Similarly, the venture`s legal form, its technological intensity, profitability, and the revenues generated have no influence on how this firm was founded.

From those factors that do influence the founding sequences of new ventures, the strongest explanation is the founding *team*. Whether entrepreneurs start a business on their own or together with others has a strong impact on how they proceed (Pearson`s $Chi^2$ = 56.032; Cramers` V = 0.158). This result is particularly interesting as it partly explains the difference between the founding sequences of clusters 1 and 7. While cluster 1 is significantly different from the other clusters in that it contains numerous solo entrepreneurs, cluster 7 includes a high share of team entrepreneurs. For cluster 7, this indicates that the massive investments made by owners do not come only from one, but rather from several entrepreneurs.

For cluster 1, it is furthermore interesting to note that it contains a high share of necessity entrepreneurs. As indicated in table 3, *start-up motives* of entrepreneurs constitute another statistically significant factor explaining differences in venture creation sequences (Pearson`s $Chi^2$ = 21.149; Cramers` V = 0.148). These results suggest that people in necessity situations tend to set up a business on their own – possibly because this is the only possibility for them to earn money in the short run. Contrary to cluster 1, cluster 4 distinguishes itself from the other clusters in that it contains a significantly higher share of opportunity entrepreneurs. Note, however, that these results do not say anything about the success of ventures created by solo necessity entrepreneurs in cluster 1 compared to the opportunity entrepreneurs of cluster 4 – even though one might suspect that the latter are more successful than the former.

Further factors that influence venture creation sequences are a venture`s *industry* (Pearson`s Chi^2 = 40.056; Cramers` V = 0.094), whereby cluster 1 contains a particularly high share of entrepreneurs active in retail businesses, whereas cluster 9 contains a high amount of entrepreneurs in the agricultural sector. Cluster 2, in contrast, is characterized by a significant share of entrepreneurs active in the arts, entertainment, and recreation industries. Similarly, the *region* in which a venture is started influences the creation sequence (Pearson`s $Chi^2$ = 66.774; Cramers` V = 0.099). Interestingly, entrepreneurs in New England are particularly likely to start a business according to the sequence of cluster 4, whereas entrepreneurs in the South-East Centre of the US rather choose the founding sequence of cluster 5. Finally, the *venture`s origin* has a statistically significant, albeit small, impact on the founding sequence chosen (Pearson`s Chi^2 = 16.772; Cramers` V = 0.122).

[9] It should be noted that we did not use all 16 but only 11 clusters as a basis for these cross-tab analyses, because clusters 8, 11, 13, 14, 15 and 16 alike included only very few cases, namely 24 sequences in total. We therefore merged these six clusters into one residual cluster, which left us with an overall number of 11 clusters.

5. CONCLUSION AND DISCUSSION

What have we learned from the SA analyses of venture creation processes conducted in this paper? In short, we gained two major insights. First, we found that SA are a useful tool for the analyses of social science data. By running SA of STS sequences including event gaps, we found that the comparison of sequences with the LCS algorithm and their PAM clustering allowed us to identify 16 different ways in which entrepreneurs set-up new ventures. Since this paper focuses on methodological considerations rather than the generation of new empirical insights, we only presented some particularly insightful results on different venture creation processes. They include the finding that necessity entrepreneurs often start a venture on their own by investing personal funds for a short time period with the aim of opening a retail business. Contrary to that, opportunity entrepreneurs typically begin with investing own funds for a short time period as well, but then immediately proceed to the development of new products or services. Another group of entrepreneurs, in turn, start ventures with writing business plans so as to acquire external investment.

Theoretically, these insights are highly relevant in two respects. First, they cast doubt on mainstream entrepreneurship research of venture creation processes (see, for example, Carter, Stearns et al. 1994; Gatewood, Shaver et al. 1995; Carter, Gartner et al. 1996; Gartner, Shaver et al. 2004; Reynolds and Curtin 2008; Whittaker 2009). Assessing the PSED2 data with traditional statistical methods, many leading entrepreneurship scholars do not find empirical support for the idea that venture creation processes follow systematic patterns. These insights even lead some entrepreneurship researchers to argue that venture creation is a random process (see Vesper 1980; Reynolds and Miller 1992). Our analyses suggest that this argument is flawed as it seems to result from the use of inappropriate statistical tools. Second, our study shows that SA are a useful tool for the analyses of social-science data. This insight is of theoretical relevance to the extent that SA became a widely accepted method in the natural sciences once it had been successfully used for decoding the human genome. Social scientists have, however, remained ambiguous about its usefulness. Yet, if we define the usefulness of an analytical method by the superior insights it generates compared to other methods, then the insights into venture creation processes we generated with the help of SA demonstrate that this method is a useful tool for the assessment of social-science data as well.

In addition to the usefulness of SA, we also discussed how to use this tool. As we have noted in the beginning of this paper, critics of SA are numerous and express multifaceted doubts about the `best` way of running SA (for two particularly dedicated critics, see Levine 2000; Wu 2000). We argue that the criticism of SA mostly seems to result from the confusing variety of options that researchers have when conducting SA. Should event sequences be reported in an STS or a DSS format – including or excluding missing values? Should the distance between sequences be calculated with the OM or the LCS algorithm? How should substitution and indel costs be set? And should sequences be clustered with the Ward or the PAM algorithm? These are only the most essential questions which researchers need to address in order to conduct SA. It is thus hardly surprising that many find SA confusing rather than helpful. Yet, we used these questions to develop a decision tree which makes the alternatives of conducing SA explicit. Importantly, this decision tree is to be used in a flexible way. Researchers can cut or add new branches depending on the options they believe to deliver most promising results. But since the most meaningful way of conducting SA is usually only identified *ex post* through cluster validation, our advice for using the SA decision tree is straight-forward. In case of doubt, always do it both ways!

REFERENCES

Abbott, A. and J. Forrest (1986). "Optimal matching methods for historical sequences." Journal of Interdisciplinary History: 471-494.

Abbott, A. and A. Hrycak (1990). "Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers." American journal of sociology **96**(1): 144.

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

Audretsch, D. B. and R. Thurik (2001). Linking Entrepreneurship to Growth. Paris, OECD Science, Technology and Industry Working Papers, 2001/2.

Brzinsky-Fay, C. (2007). "Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe." European Sociological Review **23**(4): 409–422.

Carree, M. and R. Thurik (2003). The Impact of Entrepreneurship on Economic Growth. Handbook of Entrepreneurship Research: An Interdisciplinary Survey and Introduction. D. B. Audretsch and Z. J. Acs. New York, Springer**:** 437-471.

Carter, N. M., W. B. Gartner, et al. (1996). "Exploring Start-up Event Sequences." Journal of Business Venturing **11**(3): 151-166.

Carter, N. M., T. M. Stearns, et al. (1994). "New venture strategies: Theory development with an empirical base." Strategic Management Journal **15**(1): 21-41.

Cheng, Y.-T. and A. H. Van de Ven (1996). "Learning the Innovation Journey: Order out of Chaos?" Organization Science **7**(6): 593-614.

Curtin, R. (2009). Panel Study of Entrepreneurial Dynamics II - Codebook, Survey Research Center - University of Michigan.

Davidsson, P. (2008). The Entrepreneurship Research Challenge. Cheltenham, Edward Elgar Publishing.

Dijkstra, W. and T. Taris (1995). "Measuring the Agreement between Sequences." Sociological Methods & Research **24**(2): 214-231.

Elzinga, C. H. (2003). "Sequence Similarity: A Nonaligning Technique." Sociological Methods & Research **32**(1): 3-29.

Elzinga, C. H. (2007). Sequence Analysis: Metric Representations of Categorical Time Series. Amsterdam, Vrije Universiteit Amsterdam; available at http://home.fsw.vu.nl/ch.elzinga/MetricsRevision.pdf.

European Commission (1999). Action Plan to Promote Entrepreneurship and Competitiveness. Luxembourg, Commission of the European Communities, available at: http://ec.europa.eu/enterprise/enterprise_policy/best/doc/actionplan_en.pdf.

European Commission (2003). Green Paper: Entrepreneurship in Europe. Brussels, Commission of the European Communities, available at http://europa.eu.int/comm/enterprise/entrepreneurship/green_paper/green_paper_final_en.pdf.

Gabadinho, A., G. Ritschard, et al. (2008). "Mining sequence data in R with the TraMineR package: A user's guide." from http://mephisto.unige.ch/traminer.

Garnsey, E., E. Stam, et al. (2006). "New firm growth: exploring processes and paths." Industry & Innovation **13**(1): 1-20.

Gartner, W. B. (1988). "Who is an entrepreneur? Is the wrong question." American journal of small business **12**(4): 11-32.

Gartner, W. B., N. M. Carter, et al. (2004). Business Start-Up Activities. Handbook of Entrepreneurial Dynamics: The Process of Business Creation. W. B. Gartner, K. G. Shaver, N. M. Carter and P. D. Reynolds. Thousand Oaks, SAGE Publications**:** 285-298.

Gartner, W. B., K. G. Shaver, et al. (2004). Handbook of Entrepreneurial Dynamics: The Process of Business Creation. Thousand Oaks, SAGE Publications.

Gatewood, E. J., K. G. Shaver, et al. (1995). "A longitudinal study of cognitive factors influencing start-up behaviors and success at venture creation." Journal of Business Venturing **10**(5): 371-391.

Hills, G. E. and R. P. Singh (2004). Opportunity Recognition. Handbook of Entrepreneurial Dynamics: The Process of Business Creation. W. B. Gartner, K. G. Shaver, N. M. Carter and P. D. Reynolds. Thousand Oaks, SAGE Publications**:** 259-272.

Jain, A. K., M. N. Murty, et al. (1999). "Data Clustering: A Review." ACM Computing Surveys **31**(3): 264-323.

Kaufman, L. and P. J. Rousseeuw (1990). Finding Groups in Data. An Introduction to Cluster Analysis. New York, Wiley.

Keister, L. A. (2004). "Capital structure in transition: The transformation of financial strategies in China's emerging economy." Organization Science **15**(2): 145-158.

Lesnard, L. (2006). Optimal Matching and Social Sciences. Working Paper n° 2006-01, Institut National de la Statistique et des Etudes Economiques.

Levine, J. H. (2000). "But What Have You Done for Us Lately? Commentary on Abbott and Tsay." Sociological Methods & Research **29**(1): 34-40.

Lissoni, F., B. Sanditov, et al. (2006). The Keins Database on Academic Inventors: Methodology and Contents. Working Paper no. 181, CESPRI Università Bocconi (Milan).

Low, M. B. and I. C. MacMillan (1988). "1988."Entrepreneurship: Past Research and Future Challenges,"." Journal of Management **2**: 139-161.

MacIndoe, H. and A. Abbott (2004). "Sequence analysis and optimal matching techniques for social science data." Handbook of Data Analysis. London/Thousand Oaks/New Delhi: Sage Publications: 387-406.

Massoni, S., M. Olteanu, et al. (2009). Career-Path Analysis Using Optimal Matching and Self-Organizing Maps, Springer.

Milligan, G. W. and M. C. Cooper (1985). "An examination of procedures for determining the number of clusters in a data set." Psychometrika **50**(2): 159-179.

Muller, N. S., A. Gabadinho, et al. (2008). "Extracting knowledge from life courses: Clustering and visualization." Lecture Notes in Computer Science **5182**: 176-185.

Prinzie, A. and D. Van den Poel (2006). Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM. Working Paper no. 2005/292, Universiteit Gent: Faculteit Economie en Bedrijfskunde.

Reynolds, P. and B. Miller (1992). "New firm gestation: Conception, birth, and implications for research." Journal of Business Venturing **7**(5): 405-417.

Reynolds, P. D. and R. T. Curtin (2008). "Business Creation in the United States: Panel Study of Entrepreneurial Dynamics II Initial Assessment." Foundations and Trends® in Entrepreneurship **4**(3): 155-307.

Sankoff, D. and J. B. Kruskal (1983). Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Reading, Mass., Addison Wesley.

Sarasvathy, S. D. (2001). "Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency." The Academy of Management Review **26**(2): 243-263.

Savova, G., T. Therneau, et al. (2006). "Cluster Stopping Rules for Word Sense Discrimination." Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together: 9.

Stel, A., M. Carree, et al. (2005). "The effect of entrepreneurial activity on national economic growth." Small Business Economics **24**(3): 311-321.
Ucbasaran, D., P. Westhead, et al. (2001). "The Focus of Entrepreneurial Research: Contextual and Process Issues." Entrepreneurship: Theory and Practice **25**(4): 57-81.
Vesper, K. H. (1980). Chapter 4: Sequences in Startup. New Venture Strategies. K. H. Vesper. Englewood Cliffs, NJ, Prentice-Hall 86-114.
Ward Jr, J. H. (1963). "Hierarchical grouping to optimize an objective function." Journal of the American Statistical Association **58**(301): 236-244.
Whittaker, D. H. (2009). Comparative Entrepreneurship: The UK, Japan, and the Shadow of Silicon Valley. Oxford, Oxford University Press.
Wu, L. L. (2000). "Some Comments on" Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect"." Sociological methods and research **29**(1): 41-64.
Zhao, Y. and G. Karypis (2002). Evaluation of Hierarchical Clustering Algorithms for Document Datasets, ACM New York, NY, USA.
Zhao, Y. and G. Karypis (2005). "Hierarchical Clustering Algorithms for Document Datasets." Data Mining and Knowledge Discovery **10**(2): 141-168.

CONTACT: Assistant Professor | Department of Innovation and Environmental Studies | Utrecht University | Heidelberglaan 2, 3584 CS Utrecht | + 31 30 253 7462 | A.M.Herrmann@uu.nl | www.andrea-herrmann.com