Intentional Process Modeling of Statistical Analysis Methods

Charlotte Hug, Rebecca Deneckère, Ammar Aymen

Intentional Process Modeling of Statistical Analysis Methods

Charlotte Hug, Université Paris 1 Panthéon-Sorbonne, Centre de Recherche en Informatique, 90 rue de Tolbiac, 75013 Paris, France, charlotte.hug@univ-paris1.fr (Corresponding author)

Rebecca Deneckère, Université Paris 1 Panthéon-Sorbonne, Centre de Recherche en Informatique, 90 rue de Tolbiac, 75013 Paris, France, rebecca.deneckere@univ-paris1.fr

Ammar Aymen, Université Paris 1 Panthéon-Sorbonne, Centre de Recherche en Informatique, 90 rue de Tolbiac, 75013 Paris, France, ammar.aymen@gmail.com

Abstract: Each Humanities researcher has its own way to deal with data (collection, coding, analysis and interpretation). All these specific ways of working are not shared - each researcher is reinventing his/her own method while analyzing data without any previous experience. Nevertheless, developing and sharing these methods should be useful to the research community and students in Humanities. Moreover, a lot of data analysis is done with statistical analysis methods, to find correlations between events, to make predictions or assumptions on facts or artifacts; and the use of one method or another requires good statistical knowledge. We conducted interviews among archaeologists and historians to understand their ways of working and collected information on the methods they used. We then built a method to guide researchers in using statistical analysis methods.

Keywords: statistical analysis methods, method, intentional process.

1. Introduction

Humanities include the sciences having for objective to study the human beings: their actions, their relationships as well as their traces (human cultures, lifestyles, social behaviors, societies). Studying the human beings lifestyles and their social behaviors means looking for and analyzing data that can be described under various formats or written in foreign languages, whether in the form of texts or drawings. This data can be large-sized and stored under various forms. The development of the computing tools allowed the preservation and the exploitation of big masses of data and numerous statistical methods were designed since the beginning of the 20th century to analyze them.

Statistical analysis methods are widely used in Humanities in order to find correlations between events, to make predictions or assumptions on facts or artifacts (Canning 2014). However, the use of statistical analysis methods requires previous knowledge in statistics, selecting one method or another depends on the type of data, the project's criteria, and the objectives of the researcher. The selection of a method has then a great impact on the obtained results, therefore on the interpretation. Using an unsuitable method can then lead to misinterpretations which have to be avoided at all cost for research sake. Statistical tools as R (R 2014) or SAS (SAS 2014) allow to manipulate data and to run statistical methods but researchers need to be guided during this process and during the interpretation of the obtained results, especially if they are beginners. Moreover, humanities researchers have their own ways to deal with data: the collection, coding, analysis and interpretation are often specific to each researcher. These methods are not shared - each researcher is reinventing the wheel when he/she analyzes data without previous experience. Nevertheless, modeling and sharing these methods could be useful to the research community and students in Humanities, but it is difficult to formalize them in an understandable way.

Our objective is to provide guidance to any humanities actor, from students to researchers while conducting data analysis. The provided guidance must fit the situation at hand: each project has its own constraints and specificities. It is necessary to provide a method that is flexible to better guide and support the actors.

In this paper, we present a method to guide researchers to use statistical analysis methods. This method is described as a flexible process based on the intentions of the researcher. This paper is organized as follows, section 2 presents the followed research methodology, section 3 briefly introduces the modelling language we used to represent the method and the process models of the statistical analysis methods, section 4 presents the proposed method and section 5 concludes the paper.

2. Research methodology

We first conducted a literature review on the main statistical methods used in Humanities. We found that the Principal Component Analysis (Pearson 1901) (Hotelling 1933), the Correspondence analysis (Benzecri, 1982), the Multiple Correspondence Analysis (Benzécri 1973), the Hierarchical Clustering (Sokal & Sneath 1963) and the Logistic Regression (Berkson 1944) were the most used statistical analysis methods. While we conducted the literature review and studied the different methods, we defined in parallel the corresponding process models using Map (Rolland, Prakash & Benjamen 1999), an intentional process modeling language (see section 3).

We then conducted interviews with researchers of the university Paris 1 Panthéon-Sorbonne from different fields: anthropology (Subject 1), medieval history (Subject 2), social history (Subject 3) and statistics. The objectives of these interviews were: to understand and analyze the statistical methods Humanities researchers use, to understand what were the methods they use to collect data, to understand their ways of working and the projects they were working on and to establish a generic method to guide the use of statistical analysis methods. We built a questionnaire comprising 12 questions; the interviews lasted around 2 hours. Table 1 presents the questions asked to the researchers.

Q1: could you present yourself (function and research domain)?
Q2: what are the projects you are working on?
Q3: what is the followed process to carry out a project?
Q4: how do you formulate a research hypothesis?
Q5: what are the methods you use to collect data ? Where and how do you collect data? How do you know which data to collect?
Q6: what are the statistical analysis methods you frequently use?
Q7: upon which criteria do you select one statistical method to analyze your data?
Q8: upon which criteria do you base yourself to select the variables to analyze?
Q9: how do you interpret the results obtained by statistical analysis?
Q10: do you use other statistical analysis to confirm the results of a first analysis?
Q11: how do you take into account the incomplete data during the analysis?
Q12: could you look at this model and tell us if there are things you do or not? Do you do something else?

Table 1. The questionnaire presented to the researchers.

We present below a synthesis of the answers to questions 2 to 10 of the three subjects.

- Q1 and Q2 were introductory questions to put the subjects at ease.
- Q3: what is the followed process to carry out a project? The subjects start by defining the problematic and its interest by analyzing the sources (feasibility, reliability). Then they design a database and encode the data to do statistical analysis. They select the variables to analyze and analyze the results from the point of view of the domain (as archaeologist or historian).
- Q4: how do you formulate a research hypothesis? For Subject 2, the hypothesis is formulated from the sources, from reading on a particular topic, or by attending a seminar. For Subject 3,

it consists in determining the research space, searching for personal data and formulating the hypothesis when the data is encoded.

- Q5: what are the methods you use to collect data? Where and how do you collect data? How do you know which data to collect? Subject 1 collects two types of data: the intrinsic data which is extracted from the remains or the objects themselves and the contextual data that concerns the environment in which the object was found. Subject 2 and 3 use archives, documentary research (paper, image or multimedia), online data (collaborative tools), archaeological traces.... Subject 2 structures the information by transcription (Excel, csv…) and he selects the variables until finding the good ones. Subject 3 select the data on what interest him in relation to the research project and determine whether this data can be used or not and how.
- Q6: what are the statistical analysis methods you frequently use? The usual statistical analysis methods used by the subjects are ACP, CAH, ACM AFM, and more and more often sequential analysis (subject 2).
- Q7: upon which criteria do you select one statistical method to analyze your data? Subject 1 and 2 state the selection of the statistical method is based on the nature of the variable. The approach of Subject 3 is empirical as he tests the methods and tries to exploit the results.
- Q8: upon which criteria do you base yourself to select the variables to analyze? Subject 1 uses variables related to his research topic. Subject 2 tends to use all the variables to determine two different groups of population to apply statistical methods on the group where the data is complete and the other one where it will be possible to determine the hypothesis. Subject 3 first validates the variables through khi-2 test and through his knowledge of the topic.
- Q9: how do you interpret the results obtained by statistical analysis? Subject 1 uses the chronology given by the results to interpret them. Subject 2 and 3 state that they do not use any particular methods to interpret the results of statistical analyses.
- Q10: do you use other statistical analysis to confirm the results of a first analysis? Other statistical analysis methods can be used to confirm and balance the obtained results.
- Q11: how do you take into account the incomplete data during the analysis? Subject 1, as an archeologist, tends to use samples that are carefully selected and classified to be the best as possible. Subject 2 analyses the incomplete data separately and Subject 3 explains that some colleagues use smoothing techniques to deal with incomplete or incoherent data, but he prefers to isolate them.
- Q12: could you look at this model and tell us if there are things you do or not? Do you do something else? All the subjects recognized their ways of working in the presented models. The main intentions were represented and the strategies to achieve them were adequate. However, the defined terminology was not always clear.

These interviews then helped us to better understand the researcher problems while analyzing data. The interviews also helped us to better formalize the proposed method by using the adequate terminology and by providing more detailed guidance. They also confirmed that the selected statistical methods were relevant. We gave the researchers feedback after the interviews and they were satisfied with the final version of the method.

We then built the final method as a method family that consists in describing “a set of several organized method components for a specific domain” (Kornyshova, Deneckère & Rolland 2011).

3. State of the art

3.1. Map: Intentional process modelling language

The modeling language we used to formalize the method is called Map (Rolland, Prakash & Benjamen 1999). This language allows representing processes focusing on the intentions to achieve and on the strategies to adopt to reach these intentions. Figure 1 presents a simple example of a map process model where one intention is defined as a node *Visit Paris*. A map process model always begins with a

start intention and ends with a *stop* intention. The strategies are defined as edges between nodes. In this example there are 3 strategies to achieve the intention *Visit Paris*: *By visiting websites*, *By traveling for leisure* or *By attending CAA 2014* (which took place in Paris). When the intention *Visit Paris* is achieved, the process can stop *By satisfaction* of the intention if the visit was good, or *by discontent* if the person was not satisfied of the visit.

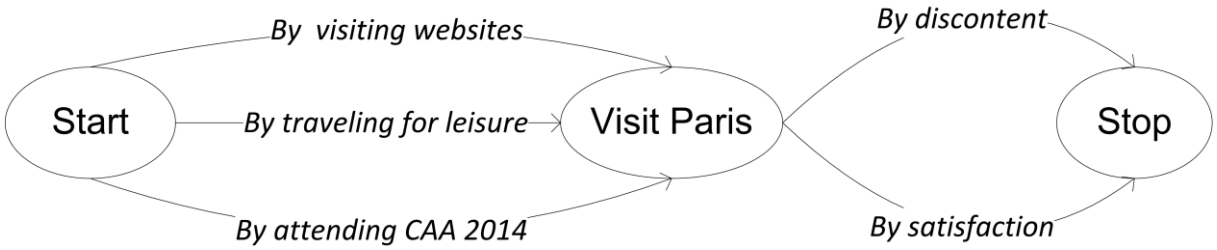

Figure 1. Example of a map process model.

Map modeling language allows representing flexible processes as several strategies can be applied to achieve a given intention. For instance, if the strategies *By visiting websites* is not sufficient to achieve the intention *Visit Paris*, it is possible to apply other strategies, as long as necessary. On the contrary to activity oriented process models, Map models do not have to be enacted sequentially: one can enact the process as long as the intentions are not achieved without following a particular order.

Map modelling language has proved to be efficient and useful in a lot of different domains: to study strategic alignment (Thévenet 2007) (Rolland 2004), to specify the outcome of business process models (Salinesi 2003), to support guidance (Rolland 1993), (Deneckère 2010), to describe intentional services (Rolland 2010), to express pervasive information systems (Najar 2011), to define systems requirements (Ralyté 1999), to study users' behavior to identify and name use cases, to tailor methods (Ralyté 2003), and also in Humanities to model scientific processes (Hug, Salinesi, Deneckère & Lamassé 2012).

We believe the Map process modelling language is adapted to represent the process of using statistical analysis methods as it is flexible and close to the human ways of thinking and working.

3.2. Statistical analysis methods

In this section we will briefly describe the selected statistical analysis methods and their corresponding process model represented with Map. Figure 2 presents the five different methods with their specific intentions and strategies.

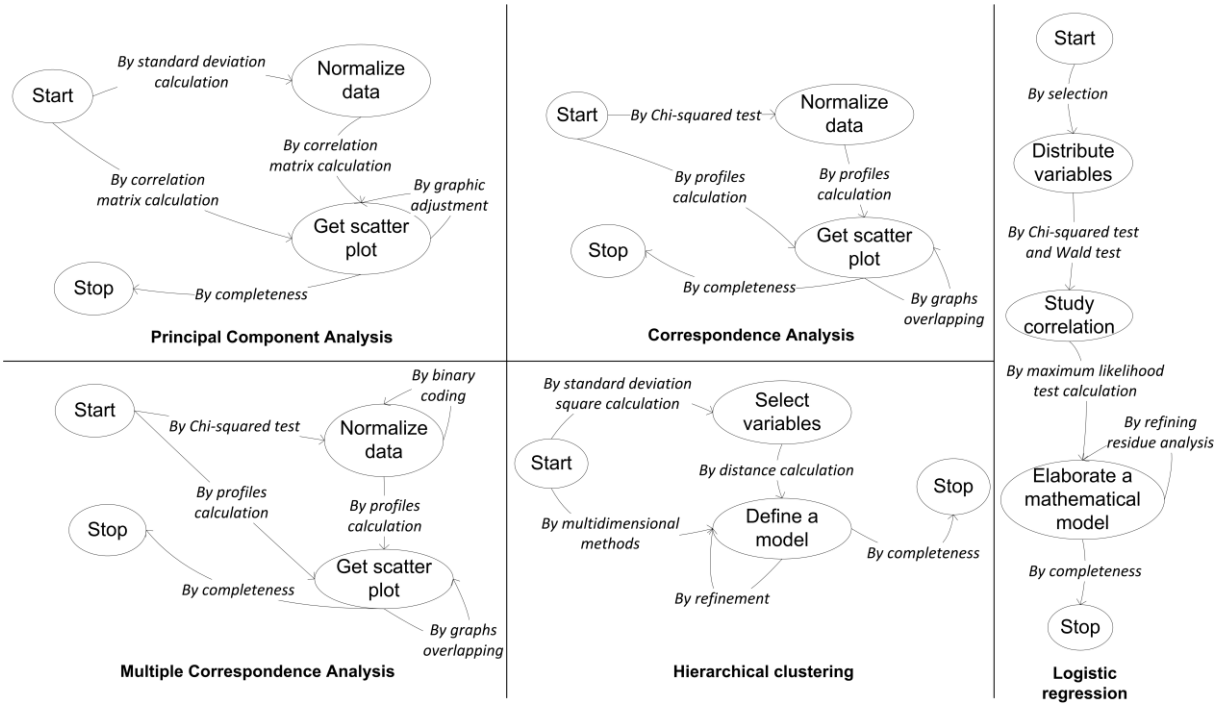

Figure 2. The maps of the different statistical analysis methods.

The Principal Component Analysis (PCA) (Pearson 1901) (Hotelling 1933) is a technique of factor analysis. The analysis takes as input a table with n rows and p columns and aims to reduce the size of the table by determining new reduced variables containing more information. This method provides a scatter plot (with adjustment of the cloud computing individuals and inertia calculation to measure the dispersion of the cloud using a correlation matrix).

The Correspondence Analysis (Benzecri, 1982) is similar to PCA but while the PCA applies to continuous numeric variables, the CA applies to two categorical variables. The method is used to study the link between two qualitative variables. The data normalization is performed using a Chi-square test (particularly by correlation analysis). A contingency table is established as a homogeneous comprehensive table. The average profiles (centroids) allow determining whether and how a class of individuals differs from the general population.

The Multiple Correspondence Analysis (MCA) (Benzécri 1973) is similar to a CA applied to more than two variables. The data normalization table is performed so as to obtain binary numbers. The rows represent the individuals, whereas the columns represent the variables. This table is then transformed into a complete disjunctive table. The rows still represent individuals, but the columns represent the terms (each term is connected to a variable). Once the number of variables increases, the data is represented as a hyper-contingency table (called Burt table). The similarity between individuals is determined by the number of terms in common. Thus, two terms are similar if they are present or absent in many individuals. The profile calculation is also used to get the total of the rows and columns, as in CA. Each profile forming a cloud of points is then projected into a different space. The projection onto a single plane allows highlighting a series of orthogonal directions, to study the projections of two clouds, which allows to choose the number of projections axes and to study the values representing the inertia of each axis.

The objective of the Hierarchical Clustering (Sokal & Sneath 1963) is to classify individuals sharing similarities from a set of variables. The selection of the variables representing the individuals is done using the standard deviation. The distance between the individuals is then calculated (e.g. with the calculation of the Euclidean distance or the distance Chebyshev), then the distance between groups

(one can calculate the minimum distance, maximum distance or the distance defined by the method of Ward (Ward 1963)) to determine a dissimilarity and aggregation index. The result of a HC is a hierarchy of classes such that any class is not empty, every individual belongs to a class and each class is the union of the classes that are included. To interpret the dendrogram, a partition is selected from the class hierarchy. The cleavage is carried out at a level where the inertia between classes increases suddenly and significantly. The data of the remaining subsets are considered relevant.

The Logistic Regression (Berkson 1944) first consists in studying the distribution of different variables by calculating the conditional probabilities (Rakotomalala 2011). Two approaches are proposed to estimate this probability: based on frequencies or based on the calculation of the likelihood ratio. It is also possible to use models such as logit (Berkson 1944) or probit (Bliss 1934) to "score" each individual. Then the correlations between the independent variables and the relationship between each explanatory variable and the dependent variable (Wald test) are calculated. The logistic regression model is finally built by applying the Chi-square test of the model (the variables are they related to the dependent variable?) and the chi-square maximum likelihood (Hosmer & Lemeshow 2000) to determine if the data contradict the established relationships. The adjustment of the model is necessary as it allows to judge the quality of the final model according to the data. This adjustment can be made with the tests of residual analysis as residual deviance and Pearson. The presentation and interpretation of the models can be made by Odds ratio test and residue analysis to check the quality of the regression.

4. The Method

4.1 The global map

We then used Map to represent the method to guide researchers to use statistical analysis methods. We first defined a global map (see Figure 3), presenting the whole analysis process, from the collect of the data to the interpretation of the results. This map was built based on the literature review and on the interviews.

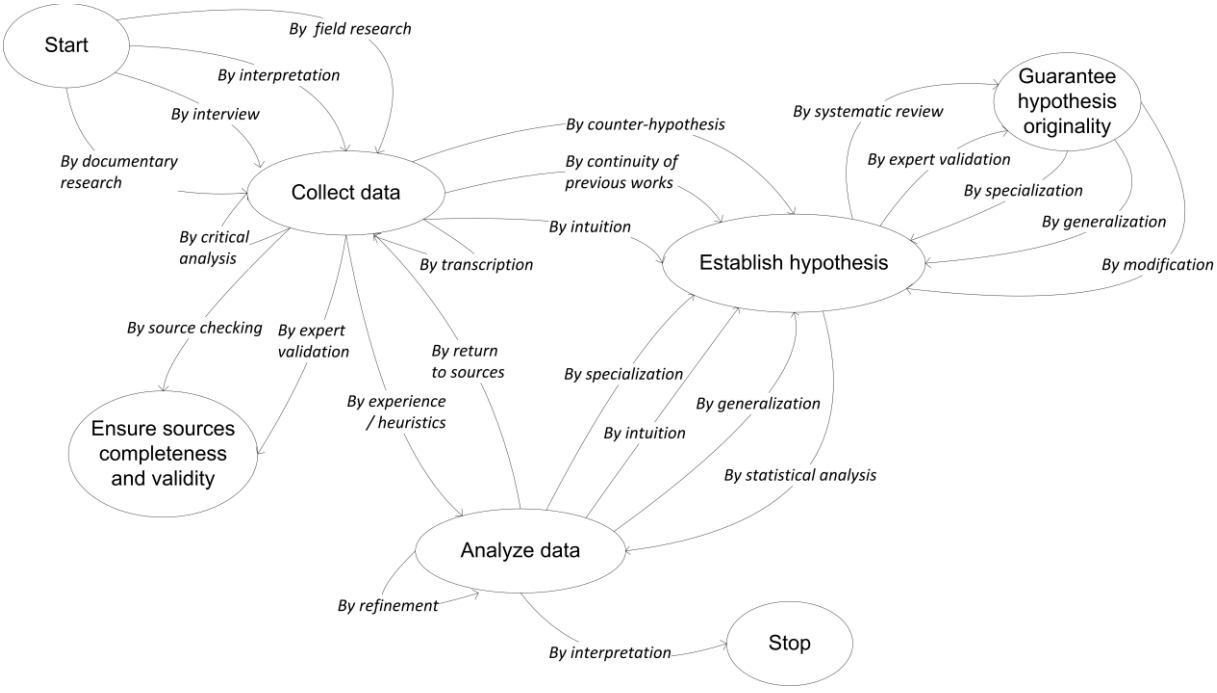

Figure 3. Global view of the proposed method.

We first defined four different strategies to *collect data*: *by interviews*, *by documentary research*, *by field research* (in the case of archaeologists for example) or *by interpreting existing works*. The researcher can enact these strategies as long as the intention *collect data* is not achieved. The researcher can also collect data *by critical analysis* of his own collect or *by transcription* when the data is entered into a computer file as a spreadsheet, a data base, an xml file, etc. If the researcher is experienced, he/she can directly *analyze data*.

During and after the process of the data collection it is important to ensure that the sources are complete and valid. We then defined the intention *Ensure sources completeness and validity* and the strategies *by source checking* and *by expert validation*. An expert of the domain will be able to detect problems (corrupt or incomplete sources), and to advise the researcher to study new potential sources.

The researcher now has the data checked. He/she can then *establish a hypothesis* that is the question he/she wants to explore according to the collected data. This can be done using 3 different strategies: *by continuity of previous works*, *by counter-hypothesis* (when a hypothesis has already been established) or *by intuition*.

Before starting the statistical analysis of the data, the researcher should verify if the hypothesis he/she wants to establish and study is original, that no researcher has already answered it in a similar way (*Guarantee hypothesis originality* intention). We then propose two strategies: *by systematic review* and *by expert validation*. If the hypothesis is weak, not original, not specific enough or on the contrary too specific, it can be changed by applying the different strategies: *by modification*, *by specialization* or *by generalization* to achieve the intention *Establish hypothesis*.

The next step is then to *analyze the data by statistical analysis*. This particular section will be detailed in section 4.2 as it is complex and requires more guidance. When the data is analyzed, the researcher can decide to change the hypothesis if the results are not satisfying: *by intuition* if he/she feels the hypothesis has to be changed, *by specialization* if it is too general, or *by generalization* if it is too specific.

When analyzing the data, the researcher can also realize the sources are not complete or should be modified; the strategy *by return to sources* is then followed. The results of the analysis can be refined to improve the visualization of the results *by refinement*.

Finally, the analyzed data can be interpreted. This is the last step of the method (*by interpretation*).

4.2 Analyze data by statistical analysis

Analyzing data by statistical analysis can be complex as there are many statistical methods and researchers need guidance while executing them. We then refine the section <Establish hypothesis, Analyze data, by statistical analysis> into another map process model. This map also comprises Start and Stop intentions. It was built from the literature review and the interviews conducted with the Humanities researchers. Figure 4 presents the map process model refining the section <Establish hypothesis, Analyze data, By statistical analysis>.

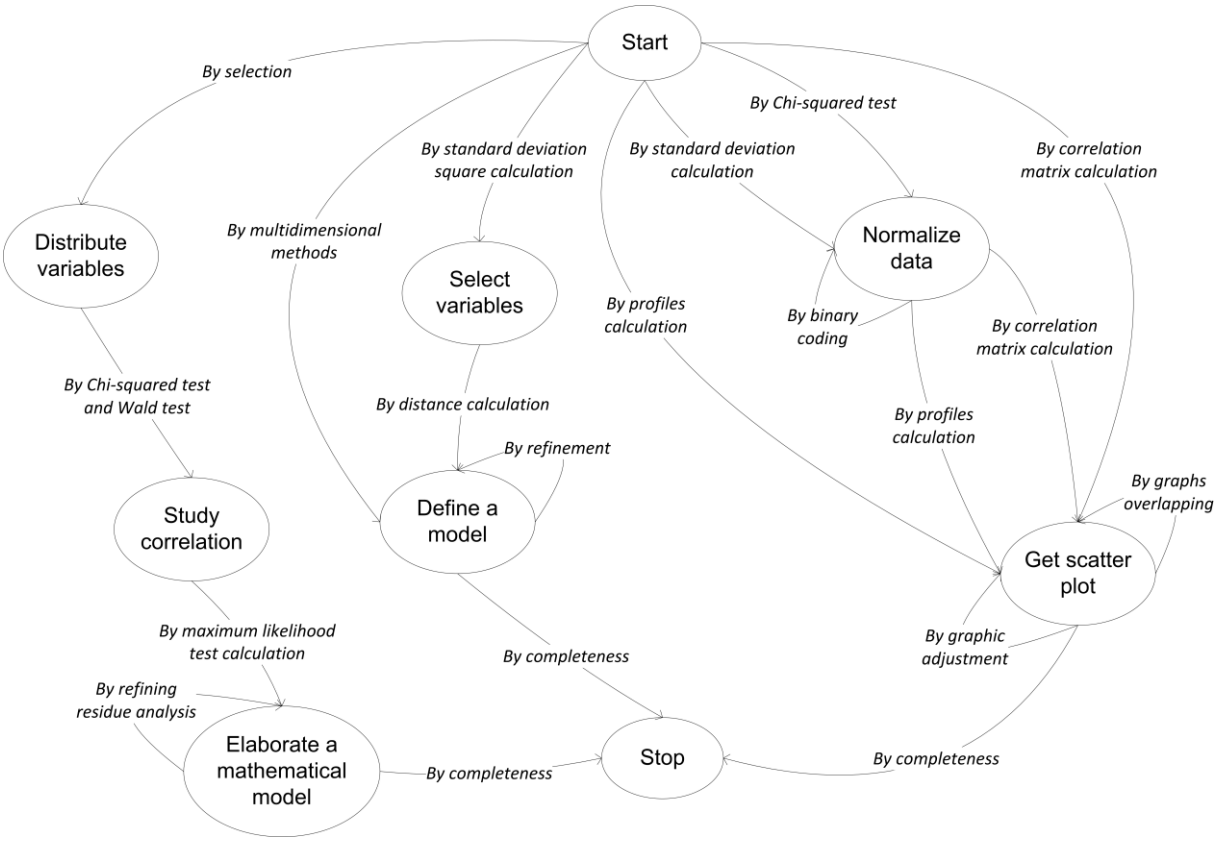

Figure 4. Refinement of the section <Establish hypothesis, Analyze data, By statistical analysis>.

The statistical methods used to construct this process model are the Principal Component Analysis, the Correspondence Analysis, the Multiple Correspondence Analysis, the Hierarchical Clustering and Logistic Regression methods. We built this map according to the models we presented in section 3.2.

Each path of this model corresponds to a particular method. For instance, the path comprising the intentions Distribute variables, study correlation, elaborate mathematical model were extracted from the Logistic regression method. The path comprising Select variables and Define a model were extracted from the Hierarchical clustering method. Some paths are common to different statistical analysis method as the intentions Normalize data and Get scatter plot that are defined both in Principal Component Analysis, Correspondence Analysis and Multiple Correspondence Analysis methods. However the strategies allow us to differentiate the methods.

5. Conclusion and future work

This work is a first step towards a tool to guide humanities researchers in using statistical analysis method. We provide a set of process models that describe the collect of data, the establishment of hypothesis and the analysis of the data. We conducted a first evaluation of the proposed models with four researchers.

We plan to validate the proposed method with master and PhD students from university Paris 1 Panthéon-Sorbonne to measure its understanding and ease of use. The experiment platform is ready and the evaluations will start in October 2014 using Google Forms and R.

The next step of this research is then to implement the method into the Online Method Engine (Vlaanderen, Spruit, Dalpiaz & Brinkkemper 2014) to provide an online tool to Humanities researchers. This tool will allow gathering all the process models for statistical analysis methods to make them available to the community, including new ones added by researchers themselves.

We also need to define the context of the data analysis projects (as the volume or type of data) to better guide the use of one statistical method or another, to help the researchers in selecting the best fitted method.

Currently, the proposed models are suitable for multi-dimensional methods, but we also plan to introduce other methods to analyze other type of data such as natural language or image. This will also be done through the Online Method Engine.

6. References

Benzécri, J.-P 1973, *L'analyse des données: L'analyse des correspondances*, Dunod, Paris.
Benzécri, J.-P 1982, *Histoire et préhistoire de l'analyse des données*, Dunod, Paris.
Berkson, J 1944, 'Application of the Logistic Function to Bio-Assay', *Journal of the American Statistical Association*, vol. 39, pp. 357-365.
Bliss, C. I. 1934, 'The method of probits', *Science*, vol. 79, no. 2037, pp. 38–39.
Canning, J 2014, 'Statistics for Humanities', Available from http://statisticsforhumanities.net [July 2014]
Deneckère, R, Kornyshova, E 2010, 'Process line configuration: An indicator-based guidance of the intentional model MAP', *Enterprise, Business-Process and Information Systems Modeling*, vol. 50, pp. 327-339, Springer, Berlin Heidelberg.
Hosmer, DW, Lemeshow, S 2000, *Applied Logistic Regression,* Wiley, New-York.
Hotelling, H 1933, 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology,* vol. 24, pp. 417-441 & 498-520.
Hug, C, Salinesi, C, Deneckère, R, Lamassé, S. 2012, 'Process modeling for Humanities: tracing and analyzing scientific processes' *Proceedings of Annual Conference of Computer Applications and Quantitative Methods in Archaeology,* pp. 245-255. Available from: Amsterdam University Press.
Kornyshova, E, Deneckère, R, Rolland, C 2001, 'Method Families Concept: Application to Decision-Making Methods, Enterprise, Business-Process and Information Systems Modeling', *Lecture Notes in Business Information Processing*, vol. 81, pp 413-427.
Mirbel, I, Ralyté, J 2006, 'Situational method engineering: combining assembly-based and roadmap-driven approaches', *Requirements Engineering*, vol. 11, no. 1, pages 58-78.
Najar, S, Kirsch-Pinheiro, M, Souveyet, C 2011, 'Towards semantic modeling of intentional pervasive information systems', I*n Proceedings of the 6th International Workshop on Enhanced Web Service Technologies*, pp.30-34, ACM, NY.
Pearson, K 1901, 'On Lines and Planes of Closest Fit to Systems of Points is Space', *Philosophical Magazine Series,* vol. 2, no. 6, pp. 559-572.
R, computer software 2014. Available from: http://www.r-project.org/
Rakotomalala, R 2011, *Pratique de la Régression Logistique*, Université Lyon 2.
Ralyté, J 1999, 'Reusing scenario based approaches in requirement engineering methods: CREWS method base', *In Proceedings of Tenth International Workshop on Database and Expert Systems Applications*, pp. 305-309, IEEE.
Ralyté, J, Deneckère, R, Rolland, C 2003, 'Towards a generic model for situational method engineering', *Lecture Notes in Computer Science*, vol. 2681, pp. 95-110.
Rolland, C 1993,'Modeling the requirements engineering process', *Proceedings of the Third European-Japanese Seminar*, Budapest, Hungary, May, pp. 85-96.
Rolland, C, Kirsch-Pinheiro, M, Souveyet, C 2010, 'An intentional approach to service engineering", *IEEE Transactions on Services Computing*, vol. 3, no. 4, pp. 292-305.
Rolland, C, Prakash, N, Benjamen, B 1999, 'A Multi-Model View of Process Modelling'. *Requirements Engineering*, vol. 4, no. 4, pp. 169-187.
Rolland, C, Salinesi, C, Etien, A 2004, 'Eliciting gaps in requirements change' *Requirements Engineering*, vol. 2, no. 1, pp. 1-15.
Salinesi, C, Rolland, C 2003, 'Fitting Business Models to System Functionality Exploring the Fitness Relationship', *Lecture Notes in Computer Science*, vol. 2681, pp. 647-664.

SAS, computer software 2014. Available from: http://www.sas.com/en_us/software/analytics.html
Sokal, RR, Sneath, PHA 1993, *Principles of Numerical Taxonomy*, W. H. Freeman and Company, San Francisco.
Thévenet, LH, Salinesi, C 2007, 'Aligning IS to organization's strategy: the INSTAL method' *Lecture Notes in Computer Science*, vol. 4495, pp. 203-217.
Vlaanderen,K, Spruit, S, Dalpiaz, F, Brinkkemper, S 2014 'Demonstration of the Online Method Engine', *In joint Proceedings of the CAiSE 2014 Forum and CAiSE 2014 Doctoral Consortium co-located with the 26th International Conference on Advanced Information Systems Engineering*, CEUR workshops, vol. 1164, pp. 169-176.
Ward, JH 1963, 'Hierarchical grouping to optimize an objective function', *Journal of the American Statistical Association*, vol. 58, pp. 236-244.