Project n°1999 / IPP/1006

"Study, development and dissemination of Data Mining tools for HLA data within the IPP"

Operational conclusions

- We built upon the observation that the former EHLASS information collection system allowed to collect data on several millions of Home and Leisure Accidents (HLA) in Europe (more than 5 000 000 observations at the end of 1999). These data were exploited statistically at national level. But it seems that the informative potential contained in the national bases -- and all the more in the recently developed (or in the process of being developed) European bases -- is widely underexploited and that it is important to better value it by using Data Mining tools and specific procedures (exploring the " mine " of data).

- Based on this observation, we led a preliminary survey with the DG SANCO and the national teams in charge of this Information System (IS). It allowed to know which software tools are used (for the statistical software packages : SAS - 5 occurences, SPSS - 3 occurences) and to notice that the already developed specific tools are rare, whereas the needs for data collection tools, data control tools and data exploitation tools are numerous and different.

- Furthermore, we explained in this survey four propositions of procedures for a better exploitation of EHLASS data :

  - **SSRD Procedure** (Synthetic Score of Relative Dangerosity) to organise into a hierarchy the potential dangerosity of products;
  - **AAS Procedure** (Automated Alarm System) to set up an automatic alert from data collected by the system;
  - **NGA Procedure** (Note of Gravity of the Accident) to define the severity of an accident;
  - **SCENAR Method** (Method of Scenarios) to define specific scenarios of accidents.

- These four procedures were the object of a marking process. When reading the marks obtained concerning decision-making utility, logical validity and the usability, they possess on average, a relatively good and comparable level of acceptability. However, the variance of the obtained notes is rather big; this indicates that some teams do not approve these methods, whereas a greater number showed an interest for these.

- We defined the Data Mining and exposed the contribution of these methods within the IPP-HLA data. Data Mining can be seen as a help process in decision making where the users themselves look for interpretation models in the data. These procedures embody statistical tools but, classic statistical methods are more descriptive and confirmative, whereas the Data Mining methods are more exploratory and decision-making oriented.

- We have further shown that the Information System concerning the HLA can be seen as one Data Warehouse where Data Mining methods are applicable. We also underlined the importance of the virtuous circle of Data Mining.

  1- To identify the data of intervention
  2- To use the Data Mining method to transform data into useful information

3- To transform information into concrete actions
4- To estimate results

- We wanted then to report on the experience that we acquired in the use of some relatively sophisticated Data Mining methods, to outline the qualities and the flaws of this type of tools. We report so experiences made with Neural Network methods and Segmentation, applied to EHLASS data.

- We noticed that the implementation of these tools is relatively heavy (recoded data, determination of the factorial axes, sharp setting of the tool's parameters, delicate interpretation) and requires some experience in the field of statistics. On the other hand, these general methods can answer only general questions. The importance and the number of specific questions stemming out of the Information System, led us to decide to finalise more adapted decision-making procedures.

- To every possible level of HLA data exploitation corresponds a range of effective tools with different characteristics. Thus, for the macro-accidentologic approach, the kind of tools which will be preferably used are of the 'indicator with epidemiological aim'-type, where the statistical quality of the tool is essential. For the micro-accidentologic approach, besides the classic statistical description tools and Data Mining tools, we suggested developing specific tools with pragmatic character where the statistical validity aspect is less essential. These tools concern the exploitation of non-aggregated data, more particularly those collected from Emergency units.

- We willingly built tools which only use variables endogeneous to the system without complex mathematical procedures. Our procedures use so only variable of the EHLASS system in additive models, from construction of scores and cross tabulation of variables. We wanted to develop simple and easy to use tools, easily understandable and running on available data as they stand.

- We exposed in details the four procedures and the developed methods, taking into account remarks and suggestions made in the survey and by the Data Mining group of experts. This group was set up with our Danish (NIPH) and Austrian (Institut Sicher Leben) partners. So, as exposed initially, rather than developing the NGA procedure, we decided to develop the SSC (Severity SCale) procedure dealing with the same subject, such as it was proposed by our Danish partners. We then developed chosen procedures by using the most widespread statistical software among the teams: the SAS (SAS INSTITUTE) software.

We made available to the teams, on the Internet:
- the SAS procedures,
- the documentations and,
- a test data file.

- Whatever the degree of interest of the developed procedures, we show more broadly in this report that it is important:

- to better value existing data by strengthening transnational cooperation ;
- to share experiences in the use of tools and software packages ;
- to use Data Mining methods in a decision-making perspective ;
- to develop, besides, other approaches and other more specific tools ;
- to test the proposed tools by taking into account the context of their development.

It is clear, however, that no Data Mining method will replace human expertise. But, human expertise can be enriched by the handling of real data, by the use of standard Data Mining methods and by the use of new tools.
Thus there can be common fertilization between the human expertise in the domain and the mastery of a set of analysis tools of which we contribute to the concrete investigation in the present report.