Department of Statistics
University of Warwick
Coventry
CV4 7AL
United Kingdom

Tuesday 29th November 2016

SHORT TERM SCIENTIFIC MISSION – FINAL REPORT OF SOPHIA WRIGHT

COST STSM Reference Number: COST-STSM-IS1304-34574
COST Action: IS1304

STSM Applicant: Sophia Wright, University of Warwick, Coventry (UK), s.k.wright@warwick.ac.uk
Host: Dr. Tina Nane, Delft University of Technology, Delft (NL), G.F.Nane@tudelft.nl
Dates of Visit: Monday 24th Oct. - Friday 4th Nov. 2016 (11 nights)
STSM Topic: Dependence Elicitation

Purpose of the STSM:

The main aim of my short term scientific mission was to enable Dr. Tina Nane and myself to continue discussions started at a COST conference held in Dubrovnik, April 2016. After discovering that we were both interested in dependence elicitation, specifically for dynamic Bayesian networks (DBNs), we agreed that it would be useful to create a formal protocol to deduce the structure of a probabilistic graphical model from experts. To our knowledge there have been numerous attempts at using experts to determine the structure of Bayesian networks (BNs), however no formal algorithm or procedure has been devised, meaning that results are not always reproducible and sometimes lack rigor.

My PhD at the University of Warwick (UK) is supervised by Professor Jim Smith and we are looking at the robustness of Bayesian networks; covering both structural misspecifications and probabilistic inaccuracies. Creating a formal protocol for dependence elicitation would therefore enhance my thesis research as well as answering an interesting question which has been left relatively unexplored in the field of expert elicitation.

As well as visiting Dr. Nane, participating in an STSM would enable me to experience the different atmosphere and research methods of a European statistics department. Although I have been fortunate to attend two large conferences this year, visiting Delft would allow me to meet other statisticians who I have not met yet and discuss ongoing research with academics who potentially share a similar field of interest to myself. The Delft University of Technology has a prestigious

reputation for using expert elicitation methods such as Cooke's Classical method in many fields, such as radiation protection and failures of gas pipelines, meaning that there is a vast array of expertise and practical experience I can learn about.

Description of the work carried out during the STSM:

During my visit to TU Delft, Dr. Nane introduced me to the work done by four previous Delft students whilst undertaking their Master's qualification. The topic that they were investigating lays down the foundations for the research we are interesting in continuing. Each Master's project built sequentially on the previous work; discussing the process of implementing learning algorithms on BNs (mostly Non-Parametric BNs) to create a partially directed acyclic graph, before turning to experts in order to determine the direction of the undirected arcs. As well as the technical information contained within these dissertations, there were also some useful literature recommendations which I had not encountered in my own initial exploration of the field.

To test our theories we have access to the Warwick Food Security BN. However, for a more self-contained and complete example, Dr. Nane introduced me to a complete dataset from her time spent at Leiden University, which is a citation database of Canadian researchers. After exploring this data and reading some of Dr. Nane's papers to see what statistical methods had already been applied to this field, we decided that it would be interesting to use a BN to forecast academic performance.

Discussing my PhD thesis with other members of the Applied Probability department in Delft generated a lot of interest from other academics who were curious to know how my approach to robustness differed from sensitivity analysis. They were also keen to discuss the ongoing application of Food Security with which my team at Warwick is currently involved. From these discussions I was given some suggestions of further work I could look into; particularly studies on the amount of information we have lost due to discretizing nodes in our Food Security BN, and the number of bins chosen during the discretization step. This STSM has therefore forwarded my thesis research by opening up new avenues to explore, as well as beginning this new work with Dr. Nane.

During my research I have been implementing BNs solely in Netica, however during this STSM visit I began looking into R packages which could learn the structure of BNs from data. As with all software, each have their own assumptions and can be restrictive or inappropriate to use depending on the dataset of interest. TU Delft has also created its own software, Uninet, specifically for BNs. Unfortunately, due to time constraints I did not have the time to use this software, however I was able to find out about any underlying assumptions and limitations that the software has so that I can make use of it in the future.

Description of the main results obtained:

To determine the structure of a Bayesian network there are three possible methods any statistician can follow. Firstly, use a learning algorithm such as the PC algorithm to fully determine the structure by first deducing collider nodes and their arcs before applying BN conditions such as ensuring there must be no cycles. Secondly, one could apply a learning algorithm to the data to find a partially directed acyclic graph, highlighting any arcs which have been left undirected. These arcs are then taken to experts and after an appropriate elicitation is concluded, the arcs are given directions. The

third option is to only use experts and not data. This option is constrained by both financial and time limits, and there is no rigorous protocol for the facilitator to follow.

Our plan is to apply each of these three methods to our Leiden citation dataset and compare the structure of the BNs produced. This would hopefully help us to determine which method is most robust. The first method of applying a learning algorithm can be done using a literature review. We have already chosen the most common learning algorithms for BNs, however further study is needed to determine which algorithm has the most suitable assumptions for our dataset. The second method of combining a learning algorithm with expert elicitation is the research already begun by Delft Master's students, however the elicitation procedures they used seemed unsatisfactory and need improvement as the results were contradictory. We believe that we have found a more statistically sound method to ask experts to direct an undirected graph by using probabilities and comparing their responses to the empirical data. The third method, of solely using experts, is the one which has the least research. We have not yet discussed in detail how we would formalise such a protocol, however it would be very interesting as there are many models which do not have data, meaning that expert elicitation is the only manner in which we can find the structure of the BN. We have started all of this prelimary work on BNs, however Dr. Nane and I intented to expand this to cover dynamic BNs is possible.

Future collaboration with the Host Institution (if applicable):

Our work from this STSM is still ongoing and we plan to keep working on the results we have obtained thus far. Once we have formalised our protocol, we would like to run an expert elicitation workshop with the citation experts at Leiden University as a working example of our theory. It is unclear at the moment whether this workshop could be held remotely, for example through Skype. However, if we deem it more appropriate to hold face-to-face meetings then another trip to Delft could be arranged in the near future.

Foreseen publications/articles resulting from the STSM (if applicable):

The research carried out during this STSM included a literature review, a brief review of BN software and their limitations, and some exploratory work into producing a dependence elicitation protocol for BNs which can be extended to cover DBNs. We have not made any precise plans to publish a paper of our work, however if we find any interesting results then it would be desirable to publish them.

Other Comments:

I would like to take this opportunity to thank the COST STSM committee for enabling me to participate in such a unique and interesting research visit. I would also like to thanks Dr. Tina Nane for suggesting we apply to the COST STSM, as well as her continued support and friendship.

Confirmation of Successful Execution (as from Tina Nane):

I hereby confirm that I have read the above report and I fully agree with its content. During Sophia's stay in Delft, we have discussed about different approaches of performing an expert judgment elicitation for learning the structure of a Bayesian network. The new ideas that emerged from our discussions and from Sophia's work in Delft can definitely be further investigated. Developing a new

protocol for eliciting the structure of a Bayesian network is not only useful for her research, as Sophia already mentioned, but also for the WG2, as well as for the entire COST community. Finally, this work is of significant importance for the field of citation analysis and bibliometrics in general, as it can show how the qualitative and quantitative aspects of evaluating researchers' performance combine.

Kind Regards,
Sophia Wright