Transport Bayesian net

Trambaiollo Luca, Esteban Muñoz Álvaro
Master's Degree in Artificial Intelligence, University of Bologna
{ luca.trambaiollo, alvaro.estebanmunoz }@studio.unibo.it

January 28, 2023

Abstract

This mini-project aims is to investigate the usage patterns of different means of transport, with a particular attention on cars and trains. Such surveys are used to assess customer satisfaction across different social groups, to evaluate public policies or for urban planning according to the new style of work after the Covid-19 pandemic mainly focusing on the pollution level in the air.
In the following report we will make use of Bayesan Network to analyse that data.
We found that a person that does the smart working has an impact in the pollution of 15% less than an other that uses personal or public transports to reach the workplace.

Introduction

Domain

Our focus is to analyze how the new possibilities to work due to the pandemic Covid-19 may have an impact to the distance and to the type of travel for different social groups and so in the pollution.
After Covid-19 many aspects of our life have changed, including certainly the way we work, with the possibility of doing smart working (unknown to many before the pandemic) or in any case doing hybrid mode. Most of the time the performance of this kind of experiments need a huge amount of data and so we also made a little experiments concerning the importance of data when learning network parameters.

Aim

The purpose of our project is to observe how much distinct kind of people with different works and style of life impact in the air pollution.

Method

We used pgmpy library methods to implement our network and run queries. These were performed to learn the dependencies between the differents nodes of the network. In addition, to understand the importance of data when learning the parameters on Bayesian network, we have generated some datasets with different number of samples and with a Bayesian estimator we have approximated the probability of the original model. Then we have compared these results with the true ones.

Results

The level of pollution is really influenced by the new style of work but not that much by the age. It is important to highlight that the amount of data needed to achieve a small enough error rate is less than what we expected.

Model

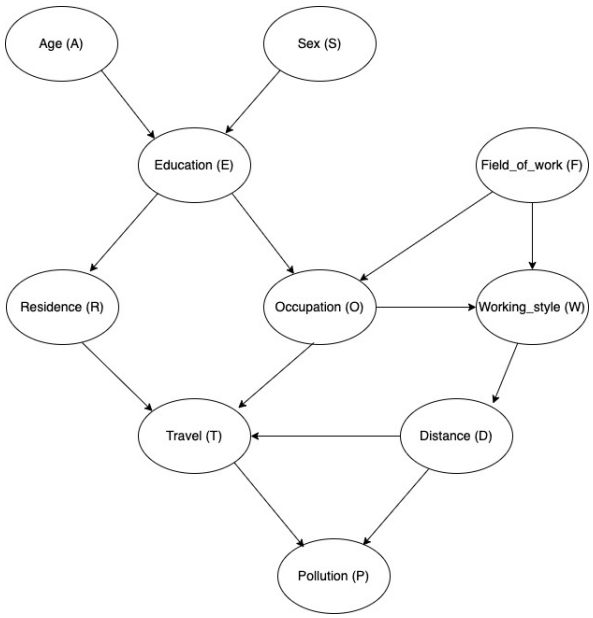

Figure 1: Bayesian network

To build the model we were inspired by the structure presented in the textbook (Marco Scutari 2015) and adding four new variables: Field of work, Working Style, Distance and Pollution to show how the new style of work impact in the travel and distance to the workplace and so in the pollution. The same work has been done regarding the probability distributions and assigning to the four new variables arbitrarily probabilities.

In the scope of this survey, each variable falls into one of these three groups: demographic indicators, socioeconomic indicators and targets. Age, Sex and Field of Work belong to the first group. In other words, they are intrinsic characteristics of the individual; they may result in different patterns of behaviour, but are not influenced by the individual himself. On the other hand, the opposite is true for Education, Occupation, Residence, Working Style. These variables are socioeconomic indicators, and describe the individual's position in society. Therefore, they provide a rough description of the individual's expected lifestyle; for example, they may characterise his spending habits or his work schedule. The last variables, Travel, Distance and Pollution, are the targets of the survey, the quantity of interest whose behaviour is under investigation.

Analysis

Experimental setup

We have performed four different groups of queries:

1. *Causal:* to compare the pollution level probabilities we have performed two queries where we have defined the pollution as query variable while the evidence variable changes between 'smart working' and 'office'.
   $P(P|W), W \in$ [sw, off]
2. *Causal:* we have generated one query for each group of age in order to understand which is the generation that has a higher probability to pollute more.
   $P(P|A), A \in$ [young, adult, old]
3. *Evidential:* to explain which is the field of work that is more conscious about the pollution we have run three queries each one with a different level of pollution.
   $P(F|P), P \in$ [high, medium, low]
4. *Intercausal:* we have defined two differents queries to justify the relation between the residence and the working style.
   $P(R|W), W \in$ [sw, off]

Another experiment has been carried out in order to analyze the consequences of data in network parameter learning: fist of all we have generated nine datasets, each one with different sizes composed by Bayesian samples. From them we have approximated the cumulative probabilities distribution (CPD) tables that we have compared to the real ones. As metric for this task we have considered the 2-norm error.

Results

The results of our queries are showed in the following section:

- *Query 1:* the probability of polluting in a high level for a person who does smart work is 29,84% while for who works in the office is 43,15%.
- *Query 2:* the difference between the probability of pollution among the generations is not high enough to assume any conclusions
- *Query 3:* we can notice that the lower the level of pollution the lower the probability of belonging to the scientific or artistic field and it happens the opposite for humanistic and social.
- *Query 4:* the difference between the probability of having a big residence being a person that does a smart working in respect to a person that works in the office is 0,8%.

From the experiment described before in the "Analysis" section we can deduce that the amount of data from where we can stop collecting samples is in the order of 10^6.

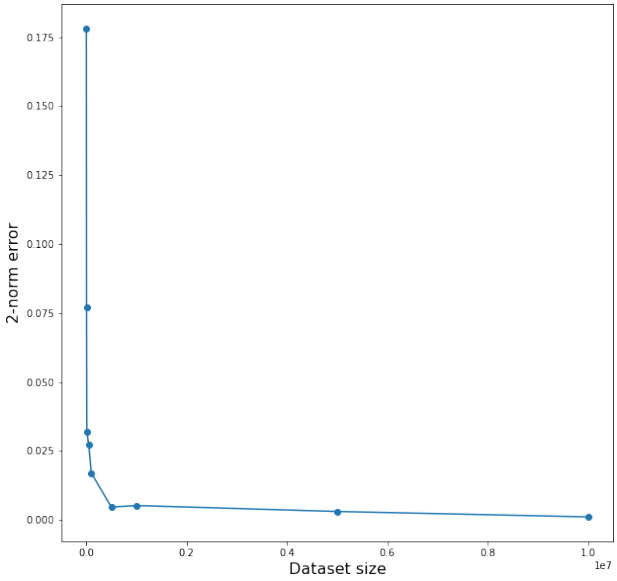

Figure 2: 2-norm error for different samples size

Conclusion

We have concluded that people who do smart working pollute less than who work in the office as well as it happens with people working in the scientific or artistic field against those who work in the humanistic or social field.
On the one hand we can underline that there is a small relevance of having a big residence between those who work in a office and those who do a smart working. Nevertheless is enough to assume that those who do smart working have a bigger residence. On the other hand, we can notice that there is not enough relevance to assume a relation between age and pollution.
At last but not least we can establish an inflection point where we could stop collecting data to train our net, in fact this point is around what we expected.

References

Marco Scutari, J.-B. D. 2015. *Bayesian Network with examples in R*. CRC Press - Taylor Francis Group.