Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics and Data Science

Master's thesis

Studying factors associated with TB positivity rate and comparing regression-based approach to Bayesian network approach.

Kedir Adem Hussen

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
dr. Stijn JASPERS

SUPERVISOR :
Dr. Sumbul HASHMI

2021
2022

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Studying factors associated with TB positivity rate and comparing regression-based approach to Bayesian network approach.

Kedir Adem Hussen

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

dr. Stijn JASPERS

SUPERVISOR :

Dr. Sumbul HASHMI

Acknowledgements

It is a pleasure to thank many people who made this thesis possible. This work would not have been possible without the support and help of them.

I would like to express my sincere gratitude to my supervisors Dr. Stijn JASPERS and Dr. Sumbul HASHMI for their insightful suggestions, motivation, dedication, continuous undiluted support on doing my thesis and encourage to finish the thesis on schedule.

I would like to thank all the professors who taught me at Hasselt University for their support to expand my knowledge all through these years. I would also like to acknowledge the VLIR-UOS that gave me an opportunity to study at Hasselt University. I wanted to express my gratitude to the EPCON company for the opportunity to work on this project.

I also extend my gratitude to all my classmates and friends at Hasselt university.

Finally, a very special thanks to my family for their endless love, continued support and encouragement, who have always stood by me in times of need.

Kedir Adem HUSSEN

June 17, 2022

Genk, Belgium

Contents

List of Tables

List of Figures

Abstract

Background: Tuberculosis (TB) remains a global public health problem. TB positivity rate is considered the golden standard for evaluating the TB reduction program. The significant geographic difference in the TB positivity rate among sub-nationals in high TB burden countries leads to investigating factors associated with TB positivity rate. Pakistan has ranked fifth among the high TB countries, which accounting for 86% of the new TB cases in 2020.

Objective: The study aimed to identify factors associated with TB positivity and to compare results from the regression-based approach with the result from the Bayesian Network approach.

Methodology: The data for this analysis obtained from a screening program that has been implemented in Pakistan. The data were a one-year aggregate at the thiessen level. The outcome of interest was the TB positivity rate for the regression approach and TB cases for the Additive Bayesian network. A negative binomial model and an Additive Bayesian network were used to address the study's objective. Environmental, socio-demographic, and access to healthcare-related factors were considered as predictors.

Results: The TB positivity rate varies across thiessen. The result revealed that the variables overall literacy rate and percentage of DPT 1 vaccine coverage were positively associated with rate of both TB positivity diagnosed either using clinical or bacteriological methods. Elevation, Distance to water features, and male-female literacy ratio, on the other hand, were all negatively associated with the rate of all forms of TB positivity. The study found that population density was negatively associated with both TB rates. Distance to water features, elevation and literacy ratio were found to be strongly related to the rate of bacteriological diagnosed TB and all forms of TB, respectively.

Conclusion: Various factors are associated with each type of TB positivity rate. Literacy rate, DPT1 vaccine coverage, and elevation can explain both rates of TB positivity in the same way. Expanding access to healthcare facilities and education would aid in identifying TB patients.
Keywords **:** *Tuberculosis (TB), TB positivity rate, Negative Binomial Model (NB), Additive Bayesian Network (ABN)*

1 Introduction

1.1 Background of the study

Tuberculosis (TB) is an infectious disease caused by the *Mycobacterium tuberculosis* bacteria [1, 2]. It primarily affects the lungs but can also affect other parts of the body [1]. Despite the availability of effective treatment, it remains a major global public health issue. TB is the second world's leading cause of death from a single infectious pathogen after COVID-19 and followed by HIV [1, 3]. The World Health Organization (WHO) reported that around 10 million people fell sick with TB globally, and an approximately 1.5 million population died from tuberculosis in 2020 [3].

The End TB Strategy, launched by the WHO, set goals to reduce global TB deaths and incidence by 95 and 90 percent in 2035, respectively [4]. It is essential to diagnose people with TB early and treat them adequately to achieve these goals. Reporting cases to a national surveillance system allows systematic and regular monitoring and treatment follow-up of TB patients. However, it was estimated that approximately 4.1 million people with TB globally were either not detected by the health system or not reported to the local/national authorities and hence missed by the national surveillance system [1]. It could be attributed to lower TB notification rates. Lower TB notifications related to less access to health facilities, limited health worker skills (or a lack of human resources), patient delays (including a lack of knowledge or fear of social/internalized stigma), patients' perceptions of TB service, and a limited TB program budget that accounts transportation, diagnosis, and treatment costs [1, 5, 6]. The missed cases could result in a long period of infection, delayed treatment, a high risk of disease complications, and eventually death. As a result, increasing the TB notification rate is one of the most important steps to reduce TB burdens and achieve the 2035 targets.

One of the possible solutions to improve low TB notification rates is by estimating the TB burden at the lower sub-national level and identifying factors associated with TB burden. There is a significant geographical difference regarding TB burdens, access to health facilities and education, and people's living standards among the sub-nationals. The sub-national level estimates would help the National TB Control Program (NTP) optimize the resource allocation to reduce the disease

burden and learn more about the barriers to notification rate[7, 8]. As a result, the rate of missed cases will decrease, while the rate of TB notification will increase.

According to the Global TB Report 2020 [1], countries with a high TB burden account for 86% of the global TB cases. Pakistan has ranked fifth among those countries, with an estimated 573,000 TB cases and 44,000 TB-related deaths in 2020. Despite the high prevalence of the disease, only about half of the cases (276,736 cases) were reported to the NTP, whereas the remaining cases were either not diagnosed or were not notified to the NTP [1, 9, 10].

The National TB Control Program (NTP) of Pakistan, which is one of many programs worldwide, is aimed at reducing tuberculosis prevalence in the general population by 2025 compared to 2012 by half and increasing notification rates [11]. The Pakistan NTP has made a significant improvement in TB notification rates, increasing by 24.45% in 2016 compared to 2012 (notified cases: 356,390 in 2016, and 269,265 in 2012)[12, 13]. Since 2016, the notification rate has been declining, and in 2020 it fell by 15.62 percent (notified cases: 327969 in 2019 [14], and 276,736 in 2020 [1]) [11]. However, the NTP wants to learn more about the factors associated with the TB positivity rate. The TB positivity rate can be related to the notification rate. The higher rate of diagnosed TB cases leads to a higher notification rate since WHO recommends that mainly all bacteriological diagnosed TB cases be reported. As a result, this research focused on identifying factors related to TB positivity rates.

The classical regression models are most commonly used in ecological studies for finding the association between the predictors and response variables. A Bayesian network (BN) can be used instead of the regression approach when the study interest is to distinguish the direct and indirect associated predictors with the response. At the same time, the Bayesian network provides the relationship between covariates. In this study, both methods are used to investigate which factor(s) are associated with the TB positivity rate.

1.2 Objective of the study

The primary goal of this study was to identify the factors associated with TB positivity rate using regression and the Bayesian network approach, specifically factors related to all forms of TB positivity rate and bacteriological diagnosed TB positivity rate. It also aimed to compare the regression approach and Bayesian network with the possibility of getting prediction.

1.3 Data Descriptions

This report was based on the data collected in Pakistan between March 1, 2021, and March 1, 2022. The screening data set consists of data collected through community screening activities conducted by a partner organization of EPCON company. The dataset contains the number of screened populations, the number of TB cases diagnosed by clinical or bacteriological methods, and the number of bacteriological diagnosed TB cases. The data were gathered at the lowest structure levels called thiessen. However, the data were only available for 300 thiessens where the community screening program was implemented. The hierarchical structure and the available data within each administrative level is presented in Table 1.

Table 1: Number of levels that had TB case record

Levels	Total	All forms of TB case N (%)	Bacteriological diagnosed TB case N (%)
Province	9	3 (33.33)	3 (33.33)
District	148	22 (14.86)	22 (14.86)
Tehsil	553	50 (9.04)	50 (9.04)
Theissen	19143	300 (1.57)	300 (1.57)

The response variables were the rate of all forms of TB positivity, denoted by R_i, and the rate of bacteriological diagnosed TB positivity, denoted by R_i^{B+}. All forms of TB cases include the cases diagnosed either using the clinical or bacteriological method. According to the WHO definition, a bacteriological diagnosed TB case is a patient with a positive culture, smear microscopy, or GeneXpert MTB/RIF. Furthermore, clinically diagnosed TB positive is a patient who has been diagnosed by a physician based on clinical examination but not bacteriological confirmed cases[15].

$$R_i = \frac{Y_i}{N_i} \qquad\qquad R_i^{B+} = \frac{Y_i^{B+}}{N_i} \tag{1}$$

where, Y_i and Y_i^{B+} be the number of all forms of TB cases and bacteriologically diagnosed TB cases in the i^{th} thiessen, respectively, and N_i be number of screened populations in the i^{th} thiessen.

The study variables included a set of variables related to environmental factor, socio-demographic factors and access to health facilities. The description of the variables along with their abbreviations used in the model is presented in the Appendix (see Table 8). The explanatory variables were chosen based on their known association with TB prevalence and public reports' availability data.

The structure of the paper is organized as follows: First, in Section 2, the methodologies used for data analysis are discussed, ranging from exploratory data analysis to statistical models. The statistical models include the regression approach, such as the Poisson and negative binomial regression models, and the Bayesian network approach, specifically the Additive Bayesian Network (ABN). This section also contains information on the model description and parameter estimation technique. In Section 3, the result of the data analysis is presented, and different models are compared. Section 4 discusses the main results, and finally, a conclusion is drawn in Section 5.

2 Methodology

This section describes the different methodologies used in this paper. The observed number of all forms of TB cases on the total screened population in particular thiessen. Similarly, the observed number of bacteriological diagnosed TB cases in a thiessen also depends on the number of observed total cases of all forms of TB cases but also the total number of screened population in the thiessen N_i. The number of screened population was different across thiessens. Therefore, the rate of TB positivity is modeled. First, Poisson regression, the most often used approach to model rate data, is discussed. Then, the negative binomial regression is introduced to account for the overdispersion issue in Poisson regression. Moreover, the data were gathe at the lowest structure level may be correlated within the next higher level; a Generalized Linear Mixed Model (GLMM) is used to account for this clustering. More details about count/ rate models and GLMM can be found in [16, 17], and [18], respectively. Finally, the Additive Bayesian Network (ABN) is introduced to model multiple variables jointly. It allows differentiating which factors are directly or indirectly associated with the response variable. More details about the ABN can be found in [19, 20].

2.1 Exploratory Data Analysis

Exploratory data analysis was used to understand better the nature and distribution of the response variables. A bar plot for the response variables was used. Moreover, a hanging rootogram was also used to check the distribution of the response variable. The rootogram plot consists of a frequency histogram (on a square-root scale) suspended from an equivalent reference distribution [17].

2.2 Generalized Linear Model

2.2.1 Poisson Regression

Poisson regression model is the most often used method for analyzing data with count or rate responses [16]. The density of the count response y_i is shown below:

$$f(y_i|X_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i}, \quad y_i = 0, 1, 2... \tag{2}$$

where μ_i is the mean of the count response, and related to the linear predictors $X'_{i\beta}$ through a link function called *log-link*. This can be written as:

$$log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} ... + \beta_k X_{ki} = X_i^{'}\beta \tag{3}$$

The *log-likelihood function* of a random sample of size n from Poisson distribution can be written in equation 4, and maximum likelihood estimation can be used to get the estimate of the regression parameters by maximize the *log-likelihood function* with a numerical iterative procedures.

$$\boldsymbol{L} = \sum_{i}^{n} \Big(y_i (X_i^{'}\beta) - \exp(X_i^{'}\beta) - ln(y_i!) \Big) \tag{4}$$

Modeling counts such as the total number of TB patients among the screened populations requires correction for the number of people who participated in the screening activity, i.e. screened population t_i. In this case, the occurrence of rate μ/t of interest and the above equation 2 become:

$$log(\mu_i/t_i) = log(\mu_i) - log(t_i) = X_i^{'}\beta \quad \Rightarrow \quad log(\mu_i) = log(t_i) + X_i^{'}\beta \tag{5}$$

where the screened population t_i is the exposure, and the log of the total screened population $log(t_i)$ is referred to as an *offset* variable, which has a fixed coefficient of one in the model. The offset variable guarantees that the predicted rate is between 0 and 1.

The primary assumption of the Poisson model is that the equality of the conditional variance of the response and its conditional mean, such that:

$$E(y_i|X_i) = Var(y_i|X_i) = \mu_i = exp(X_i^{'}\beta) \tag{6}$$

The equality between the conditional mean and condition variance for count or rate response is known as *equidispersion*. However, this assumption is often violated in practice. Hence, count data often vary more than we would expect if the response follows Poisson distribution [16]. In other words, the response's variance exceeds its mean. This condition is called *overdispersion*. The most common causes of this problem are correlated responses, excess variation between response probabilities/counts, and missing important predictor variables. Failure to account for overdispersion leads to underestimating the standard error of the estimates where one might detect the statistical significance by chance [16, 17, 21]. A negative binomial model, which is presented in the next section, can be used to handle overdispersion issue.

2.2.2 Negative Binomial Model

Negative binomial model (NB) can be used to address the issue of overdispersion since it has an additional parameter such that the variance can exceed the mean. A negative binomial distribution for a response variable Y has the following density form:

$$f(y|p,r) = \binom{y+r-1}{r-1} p^r (1-p)^y, \quad y = 0, 1, 2... \tag{7}$$

where p is the probability of success in individual trial, and r is the number of success in the trial. Hence, it can be seen that the mean and variance of Y are given by:

$$E(Y) = \mu = \frac{r(1-p)}{p} \qquad Var(Y) = \frac{r(1-p)}{p^2} = \mu + r\mu^2 \tag{8}$$

A NB model can be viewed as a generalization of the Poisson model by allowing the Poisson parameter following a gamma distribution [17, 22]. The distribution of y_i conditional to the observed covariates X_i and the unobserved heterogeneity $\tau_i = e^{\epsilon_i}$ is a Poisson with conditional mean and conditional variance $E(y_i|X_i, \tau_i) = Var(y_i|X_i, \tau_i) = \mu_i\tau_i$:

$$f(y_i|X_i, \tau_i) = \frac{exp(-\mu_i\tau_i)(\mu_i\tau_i)^{y_i}}{y_i!} \tag{9}$$

It implies that the Poisson model formulation in equation 3 can be modified as follows:

$$log(E(y_i|X_i, \tau_i)) = log(\mu_i\tau_i) = X_i\beta + \epsilon_i \tag{10}$$

where τ_i represents the unobserved manner of thiessens regarding the number of TB cases that is not fully captured by the observed covariates X_i. It follows a gamma distribution with mean 1 and variance $1/\theta$, $\tau_i \sim Gamma(\theta, \theta)$. Hence, the probability density of NB has the form of:

$$f(y_i|X_i) = \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{11}$$

This formulation of negative binomial distribution is similar to equation 7 with $r = 1/\alpha$ and $p = 1/(1+\alpha\mu_i)$, where $\alpha = 1/\theta$ is the dispersion parameter. The variance of the response equals its mean as θ approaches infinity (i.e. $\alpha \to 0$), and the negative binomial distribution converges to the Poisson distribution. Hence, a negative binomial regression model is a potential candidate to handle the overdispersion issue in Poisson regression.

Maximum likelihood estimation can also be used here to estimate the parameters of the model by maximizing the *log-likelihood function* given by:

$$L = \sum_{i=1}^{n} y_i ln\left(\frac{\alpha * \mu_i}{1+\alpha * \mu_i}\right) - \frac{1}{\alpha} * ln(1+\mu_i) + \ln\Gamma(y_i + \frac{1}{\alpha}) - ln\Gamma\left(y_i+1\right) - ln\left(\Gamma(\frac{1}{\alpha})\right) \quad (12)$$

For the rate response with an offset variable of log of the total screened population, equation 10 is written as:

$$log(E(y_i|X_i,\tau_i)) = log(\mu_i\tau_i) = X_i\beta + \epsilon_i + log(t_i) \quad (13)$$

2.2.3 Generalized Linear Mixed Model

The generalized linear mixed model (GLMM) is used when modeling data with more than one source of variability. Meaning that it can be used to analyze datasets when observations are correlated due to clustering or repeated measurements by introducing a random effect [18]. The data were collected at the lowest structural level, thiessen, which is hierarchically clustered within tehsil, district, and province, respectively. The hierarchical structure of the data is presented in Table 1. GLMMs can thus be used to analyze such data. However, after the tehsil level, the higher the structure level, the more independent the data (dissimilar). As a result, only clustering at the tehsil level is considered.

For total number of TB patients y_{ij} in the j^{th} thiessen within i^{th} tehsil, $i = 1, 2, ..n$ and $j = 1, 2, ..n_i$, where n_j is the number of thiessens within the i^{th} tehsil, the general formulation of GLMM is:

$$\eta(\mu_{ij}) = x_{ij}^{'}\beta + z_{ij}^{'}b_i \quad (14)$$

where, $\eta(.)$ is the link function of generalized linear model, $x_{ij}^{'}$ and z'_{ij} are the design matrix for fixed and random effect, respectively, β is a vector for fixed regression coefficient, and b_i is a random effect which follows a normal distribution with mean zero and variance D, $b_i \sim N(0, D)$.

Similar to the generalized linear model, maximum likelihood estimation can be used to estimate the parameters by integrating the marginal likelihood out of the random effect u_i [18], shown below:

$$L = \prod_{i=1}^{N} \int \prod_{j=1}^{n_j} f(y_{ij}|\beta, b_i) f(b_i|D) d(b_i) \quad (15)$$

where $f(y_{ij})$ and $f(b_i)$ are the probability distribution of the response variable y_{ij} and the random effect b_i, respectively. However, integrating the likelihood function of a discrete outcome analytically is difficult. As a result, numerical integration techniques are used to evaluate the integral [18].

In this study, the Poisson and the Negative Binomial models are only considered. As a result, a random effect was added to both models, as shown below, to account for data clustering within tehsil.

$$\eta(E(y_{ij}|u_i)) = log(\mu_{ij}) = x_{ij}^{'}\beta + u_i + log(t_{ij}) \tag{16}$$

where $u_i \sim N(0, \sigma^2)$ is a tehsil-specific variable that accounts for the correlation between measurements at thiessen level within the same tehsil, and $log(t_{ij})$ is the log of the total screened population considered as an offset variable.

2.3 Bayesian Network Model

Bayesian network (BN) is a type of graphical modeling that attempts to distinguish between indirect and direct associations in complex multivariate data [19, 23]. A BN consists of a set of nodes representing random variables and a set of directed links that connect pairs of nodes (often referred to as *edges*) [23]. Bayesian networks are directed acyclic graphs (DAG), which means there is no way to trace a path beginning at one node and ending at the same node following the link paths. [19, 23, 24, 25, 26]. Each random variable has a conditional probability distribution that quantifies the probabilistic relation between the node X_j and its parents, Pa_j, such that for a network of k nodes.

$$P(X) = \prod_{j=1}^{k} P(X_j|Pa_j) \tag{17}$$

As a result, the information in the network can be used to compute the full joint probability distribution. The entire nature of a set of variables' relationship can be captured by a well-represented Bayesian network.

2.3.1 Method of Bayesian Network Developing

BNs can be built using expert knowledge, or solely on datasets or a combination of the two. [26, 27]. The method of constructing BN based on expert knowledge is referred to as manual construction [27], which is mainly subjective [28]. In practice, we may not have enough information on the phenomenon, which makes it impossible to build the DAG, particularly when there are many variables. In this case, an automatic construction method is preferred, which is a data-driven construction of BN [26, 27]. This construction method required a strong assumption that the dataset had no missing observation. A combination of both methods is commonly used.

2.4 Additive Bayesian Networks

Additive Bayesian network (ABN) is a special type of Bayesian Network model that extends the standard generalized linear model to multiple dependent variables by representing the joint probability distribution of the random variables. It is a data-driven method that does not require prior expert knowledge [19]. However, it allows for the incorporation of expert knowledge by imposing specific controls on the node-node relationship, such as banning and/or retaining specific arc (i.e., the relationship between two nodes). Furthermore, unlike other BNs that used to incorporate only the same node type, ABN allows for the simultaneous modeling of different types of nodes such as Gaussian, Poisson, and Binary variables. In this study, there was insufficient prior knowledge to construct the network structure (DAG), and the response variables are count variables; thus ABN model has been chosen to analyze the data.

In the ABN approach, all the variables are modeled by specific probability distributions according to their type. In this study, all forms of TB cases, bacteriological diagnosed TB cases, and total screened populations were specified with a Poisson distribution. The remaining covariates were assumed to follow a Gaussian distribution. As there is no other way to incorporate the rates in an ABN, the denominator variable of the rates (i.e., total screened populations) was included as an additional variable by banning the relationship between this variable and other covariates. Hence, the response variables were directly modelled as a count.

In ABN, the term $P(X_j|Pa_j)$ in equation 17 can be translated using the classical notation for the exponential family parameterization and written as follows:

$$P(X_j|Pa_j) = exp(\eta(\theta_j)T(X_j, Pa_j) - A(\theta_j))H(X_j, Pa_j) \tag{18}$$

where function $\eta(.), T(.), H(.)$ are node-dependent. Thus, each node is modelled with their set of parents in a GLM approach.

2.4.1 Bayesian Learning

A Bayesian Network has two equally important components for finding and interpreting the result: a qualitative component (the structure,$\mathbf{G}$) and a quantitative component (the parameter estimates, β_A). The structural selection and parameter estimation processes are altogether referred to as *learning* [26]. The BN learning is performed as a two-step process that includes structural learning, which generates the structure of the DAG, and parameter learning, which estimates the local distribution implied by the structure of the DAG. Both learning procedures are relevant for understanding the final model and are interconnected [19, 20, 26, 29].

For an Additive Bayesian Network (ABN) model $A = (G, \beta_A)$ given that the dataset D, the model learning can be written as:

$$P(A|D) = \underbrace{P(G, \beta_A|D)}_{model\ learning} = \underbrace{P(G|D)}_{structure\ learning} * \underbrace{P(\beta_A|G, D)}_{parameter\ learning} \tag{19}$$

where $P(G|D)$ is the posterior probability of the DAG, $P(\beta_A|G, D)$ is posterior distribution of the parameters.

2.4.2 Structural Learning

The first step in BN approach is to identify the optimal model with the highest network score. The *log-marginal likelihood* is commonly used as a goodness of fit metric. The total network score, *log-marginal likelihood*, for an ABN model A is calculated as:

$$P(D|G) = \prod_{j=1}^{n} P(D_j|G) \tag{20}$$

where, D_j are the observed data at node j and G is the structure.

The optimal global model can be identified by performing an exhaustive search of the data using an exact approach method and iterating the allowed parents per node (parent limits). The number of parents allowed per node is equivalent to the number of covariates in each regression node. The maximum goodness of fit is reached when the marginal likelihood does not improve further after increasing the parent limits. In other words, the parent limits are increased until the goodness of fit remains constant, identifying the globally optimal DAG. The process of identifying an optimal model is referred to as structural learning [19, 20].

The exact search algorithm is computationally intensive for models with more than 20 variables. A heuristic search can be used instead of the exact search algorithm to find the optimal DAG. This search algorithm provides a DAG (model) close to the global optimal DAG as the number of searches increases. The heuristic search algorithm, on the other hand, does not guarantee the finding of the optimal global structure, but only of an optimal local structure [19, 26].

2.4.3 Parameter Learning

After the network structure has been identified and selected, parameter learning can be performed locally. Meaning that, only the local structure, which consists of the index node and set of parent nodes, is required [19]. In other words, a node is modeled with its parents in the GLM approach. There are two methods for estimating the parameters of a network: the maximum likelihood and the Bayesian approaches. The maximum likelihood approach assumes an unknown but fixed set of parameters that maximizes the likelihood function. The Bayesian approach used in this study assumes a random parameter and assigns a prior to them. Thus, the parameter estimate is obtained from the posterior distribution of the network using Integrated Nested Laplace approximations (INLA) inference. The estimated coefficients have the classical epidemiological interpretation depending on the index node variable: log odds ratio (odds ratio) for binary, correlation/relationship for Gaussian, and log risk ratio (rate ratio) for Poisson.

2.4.4 Adjusting for Overfitting

Once the globally optimal DAG has been identified, it is necessary to adjust for overfitting since automated procedures are frequently prone. In other words, the global optimal DAG may contain excess structure. This adjustment can be made using either parametric or non-parametric bootstrapping. The parametric bootstrap technique involves randomly simulating data of the same size as the original observed data from the global optimal DAG and then refitting it. Arcs supported by 50% of the bootstrapping are retained in the pruned (best) DAG. It was planned to use JAGS for parametric bootstrapping in this study. However, a single simulation took 2:30-3:00 hours on a standard computer. It was not carried out due to the computational time constraints of running at least 1500 bootstrapping. The code is available in the Appendix C.

2.4.5 Adjusting for Clustering

Failure to account for the data's clustering nature may result in an underestimation of variance and thus unreliable parameter estimates. As described in section 2.2.3, the typical solution to this problem is to switch from GLM to GLMM. The clustering nature is included after identifying the best DAG. This procedure is also required to perform MCMC sampling; however, it was not used in this study due to the same constraints.

2.4.6 Prior Choice

The choice of prior is critical in the Bayesian approach since it affects the posterior distribution of a parameter. Priors reflect the information that have been known to a researcher regardless of the dataset on which the model is fitted. In the context of a Bayesian network, priors are assigned to the two components of the network: the structure of the network (DAG) and the parameter network, which quantifies the effect of a parental node on a child node [26].

A uniform prior $P(G) \propto 1$ that all DAG structures were equally supported without data was chosen. A uniform prior was used to ensure that there was no preference for a structure, allowing a fully data-driven approach. It is common to use noninformative prior for regression parameters. As a result, a default noninformative prior was chosen. For all regression parameters, including the

intercept, a normal distribution with mean zero and variance 1000 has been used. All these prior are the default values in **abn** package, and there is no way to change them to perform a sensitivity analysis. The priors can be expressed as follows:

$$\beta_i \sim N(0, 1000) \tag{21}$$

for i=1,2,...,n where n is total number of additive term in the network.

2.5 Software

All the analysis were performed using the software R version *4.2.0*. The package **abn** was used to create the Bayesian network. The code is included in the Appendix C.

3 Results

3.1 Exploratory Data Analysis

Figure 1 shows the distribution for number of TB cases across thiessens. It is revealed that the cases vary across thiessens. The average number of all forms of TB cases was 3.33, and the average number of bacteriological diagnosed TB cases was 1.42, with standard deviations of 3.713 and 1.92, respectively (see Table 2). The distribution of tuberculosis cases is highly skewed, with the highest TB cases of 23 and 14 diagnosed by either clinical or bacteriological method and only by bacteriological method, respectively. During the study period, the average rate of diagnosing TB cases by clinical or bacteriological method was 627.1 per 10000 screened population. The average rate for bacteriological diagnosed tuberculosis was 272.2 per 10000 screened population, which is lower than the rate of all forms of tuberculosis. Most of the thiessen's had no record of TB cases for both seniors, accounting for 21.66 percent for all forms of TB cases and 40 percent for bacteriological diagnosed TB cases. In addition, the empirical means and variances were not equal, resulting in a ratio greater than one. The ratio of more than one could suggest the presence of data overdispersion.

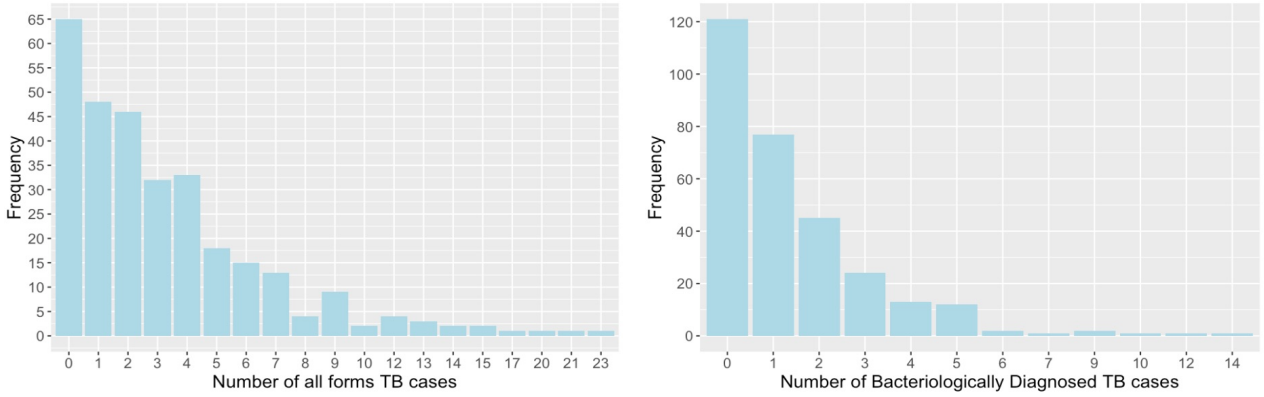

Figure 1: Distribution of TB cases record: all forms of TB (left) and bacteriological diagnosed TB cases (right)

Table 2: Empirical mean, variance, standard deviation (Std.) and variance-mean ratio of TB cases

	Mean	Variance	Std.	Variance to Mean Ratio
All forms TB cases	3.33	13.79	3.713	4.15
Bact+ TB cases	1.41	3.69	1.92	2.61

The rootogram graph for both TB positivity rates are presented in Figure 2. The hanging bar over the x-axis in the rootogram represents overestimated observation. On the other hand, the hanging below the horizontal axis represents underestimated observation. The result from Figure 2 suggests that a negative binomial model would fit the data better than Poisson regression since the negative binomial distribution is less deviated from the observations compared to the Poisson distribution. Furthermore, the best candidate distribution for the data can be selected based on the likelihood ratio test and the model-fit criteria statistic after including important variables in the model.

3.2 Results for Regression-based Approach

A Poisson regression model was fitted first. The overdispersion parameter was 2.42, which is greater than one. A boundary likelihood ratio test (LRT) was used to test the presence of overdispersion in the model since the parameter estimate can lie on the parameter boundary space under the null hypothesis $\alpha = 0$. As a result, a mixture Chi-squared distribution for the likelihood ratio test is used. The p-value of a boundary LRT is obtained by adding equally weighted p-values of the LRT with a chi-squared distribution of zero and one degree of freedom. The likelihood ratio test was 56.10 with a p-value smaller than 0.0001. The result was highly significant, and it indicated the presence of overdispersion. The result implies that, on average, the difference between the fitted and observed rate of TB positivity are considerable larger than the specified by the Poisson distribution. As a result, a negative binomial model was fitted. Moreover, The rootogram graph for Poisson and Negative binomial after accounting for important covariates in the model in Figure 6 also suggests that the negative binomial model fits the data better than the Poisson since most of the hanging bars are close to zero.

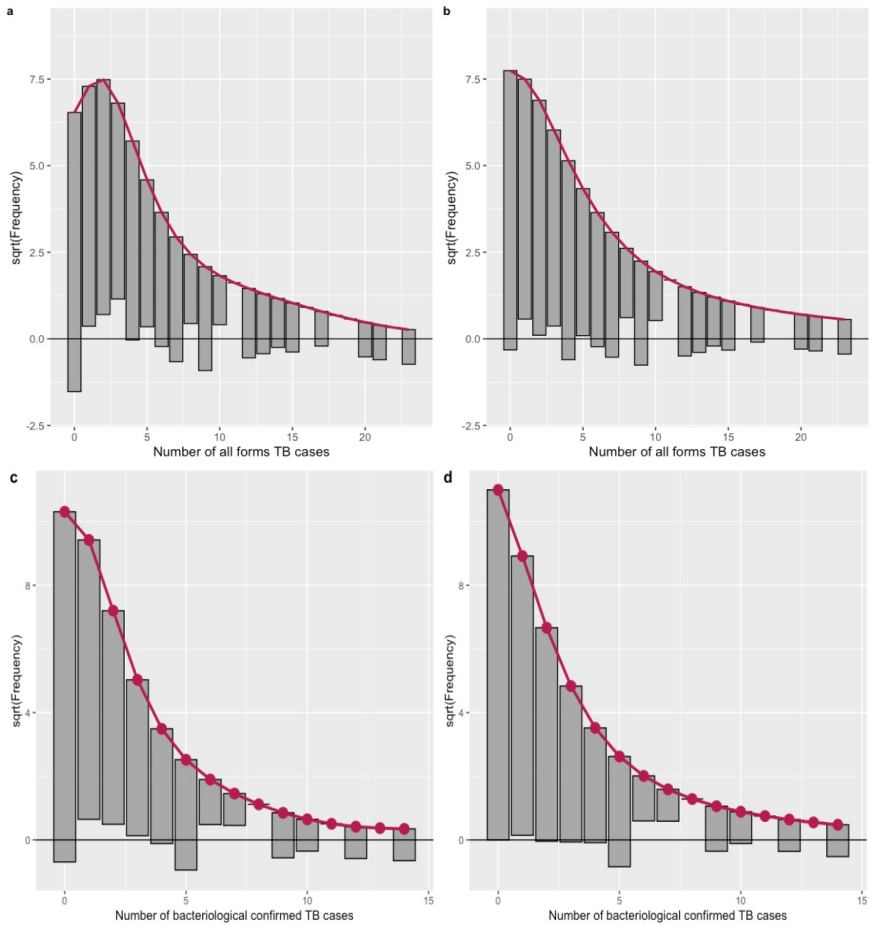

Figure 2: Distribution of TB cases: Poisson (**a** and **c**) and Negative Binomial (**b** and **d**) for all forms of TB cases (top) and bacteriological diagnosed TB cases (bottom)

Furthermore, GLMM for negative binomial was fitted to account for the clustering of observations within the same tehsil. The boundary LRT suggests that including the random effect in the negative binomial does not improve the result ($LRT = 0.583, p - value = 0.2226$). Based on model-fit criteria statistic, the standard negative binomial model has lower Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) values. It also revealed that the standard negative binomial (NB) provides a better fit than its corresponding GLMM (in Table 3). Hence, the standard negative binomial model (NB) was selected for analyzing the data.

Table 3: Model Comparison

	All forms of TB positivity		Bacteriologically diagnosed TB positivity	
	AIC	**BIC**	**AIC**	**BIC**
Poisson	1307.505	1340.839	896.0644	933.1023
Negative Binomial (NB)	1255.755	1289.089	877.4593	907.0895
GLMM Poisson	1295.443	1332.48	-	-
GLMM NB	1257.172	1294.21	-	-

Similarly, Poisson and Negative Binomial models were fitted to identify factors associated with bacteriological TB positivity rate. A boundary likelihood ratio test revealed that the dispersion parameter in the Poisson model 1.913 was highly significant (LRT:21.83, p-value:< 0.0001). The GLMM for Poisson was not fitted because the GLMM for NB did not converge, making the comparison with the GLMM Poisson impossible. Using model-fit criteria, the NB model has lower AIC and BIC than Poisson, in Table 3. Therefore, the NB model was chosen to fit the data.

Table 4 shows the estimated regression coefficients, standard error (Std.Error) and p-value for each variable from the negative binomial models. The result shows that DPT1 vaccine coverage and overall literacy rate are positively associated with both TB positivity. Distance to water features is negatively associated with TB positivity diagnosed by clinical but not significantly associated with bacteriological TB positivity rate at 0.05 level of significance. Elevation is negatively associated with both TB positivity. Additionally, and male-to-female literacy ratio is negativelyassociated with all forms of TB positivity rate, whereas poverty is positively associated with all forms of TB positivity. Measles vaccine coverage, on the other hand, is negatively associated with the rate of TB diagnosed by bacteriological method. Stunting and distance to artificial surfaces are both positively associated with bacteriological TB positivity rate.

All the variables were standardized before fitting the model since the variables are expressed on different scales, which leads to unrealistic estimates and standard error. Hence, the estimates can be interpreted as an increase or decrease in one unit standard deviation from the corresponding average values.

At the thiessen level, a one percent increase in the overall literacy rate, DPT 1 vaccine coverage, and the proportion of people living in poverty increases the rate of TB confirmed by clinical or bacteriological methods by 30%, 39%, and 16%, respectively, while other covariates remain constant. Living one kilometer away from water features increases the rate of all forms of tuberculosis positivity by 36% while controlling for other covariates. Furthermore, a one percent increase in overall literacy rate and DPT 1 vaccine coverage at theissen level increases the rate of bacteriological diagnosed tuberculosis by 90 percent and 24 percent, respectively. Moreover, a one-meter increase above the sea level is associated to a 68 percent decrease in rate of tuberculosis positivity diagnosed using the bacteriological method. A one percent increase of stunning increases the bacteriological diagnosed TB positivity rate by 25%. However, a one percent increase in the measles vaccine coverage is associated with an increase in bacteriological diagnosed TB positivity rate of 32%.

Table 4: Parameter estimate, standard error (Std.Error), and p-value for Negative Binomial model

	All forms of TB positivity rate			**Bact+ TB positivity rate**		
	Estimate	**Std.Error**	**p-value**	**Estimate**	**Std.Error**	**p-value**
(Intercept)	-3.1211	0.0467	$< 0.0001^{***}$	-4.00864	0.06497	$< 0.0001^{***}$
demog_dpt1	0.3315	0.0692	$< 0.0001^{***}$	0.64151	0.14394	$< 0.0001^{***}$
demog_literacy	0.2683	0.0726	0.0002^{***}	0.21888	0.07598	0.0039^{**}
infra_water	-0.4464	0.0980	$< 0.0001^{***}$	-0.22054	0.12453	$0.0765.$
env_elevation	-0.3690	0.0884	$< 0.0001^{***}$	-0.3752	0.1251	0.00271^{**}
demog_literacy_ inequality	-0.1266	0.0658	$0.0544.$			
demog_poverty	0.1448	0.0622	0.0199^{*}			
infra_cluster2hcf	0.0837	0.0516	0.1051			
demog_measles				-0.3524	0.11310	0.0018^{**}
demog_childst				0.22495	0.07606	0.0031^{**}
env_dst190				0.12713	0.06031	0.0350^{*}
θ	4.185	0.920	$< 0.0001^{@}$	3.88	1.22	$< 0.0001^{@}$

$***p - value < 0.001$; $**p - value < 0.01$; $*p - value < 0.05$; $.p - value < 0.1$;

@ $p - value$ based on likelihood ratio test

3.3 Results for Additive Bayesian Network Model

Here, ABN model 1 represents the additive Bayesian network for all forms of TB cases, and ABN model 2 represents the additive Bayesian network for bacteriological diagnosed TB cases.

The optimal DAG was found using an exact search by increasing the parent limits sub-sequentially from one to seven. Figure 3 shows the total network scores for various parent limit values. The log-marginal likelihood value for both models did not improve after allowing five parents per node, as shown in Figure 3. The log-marginal likelihood value for the optimal DAG of ABN model 1 and 2 are -14845.76 and -14605.94, respectively.

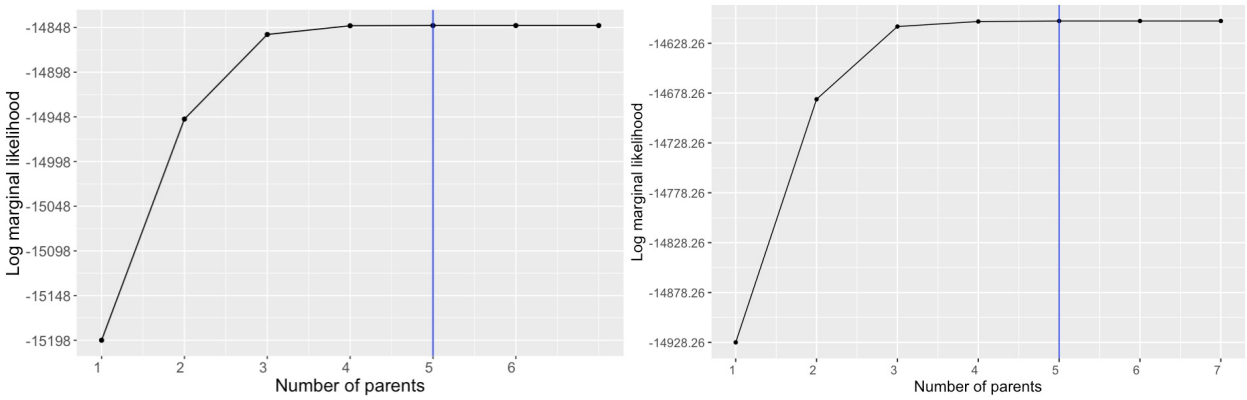

Figure 3: Network complexity for all forms TB cases **(left)** and bacteriologically diagnosed TB cases **(right)**

The optimal DAG for ABN model 1 has parent limit of 5, 20 nodes, and 44 arcs, whereas the optimal DAG for ABN model 2 has parent limit of 5, 20 nodes, and 43 arcs. Figure 4 and 5 shows the relationship among the variables in the optimal DAG of model 1 and model 2, respectively. The dashed lines represent negative association/relationship between variables in both figures, whereas the solid black lines represent positive association/relationship between variables. The diamond nodes represent Poisson distribution, while the oval nodes represent normally distributed random variables. The diamond node with yellow represents the outcome of interest. The gray oval shape with an outgoing blue line represents the factors directly associated with the outcome of interest.

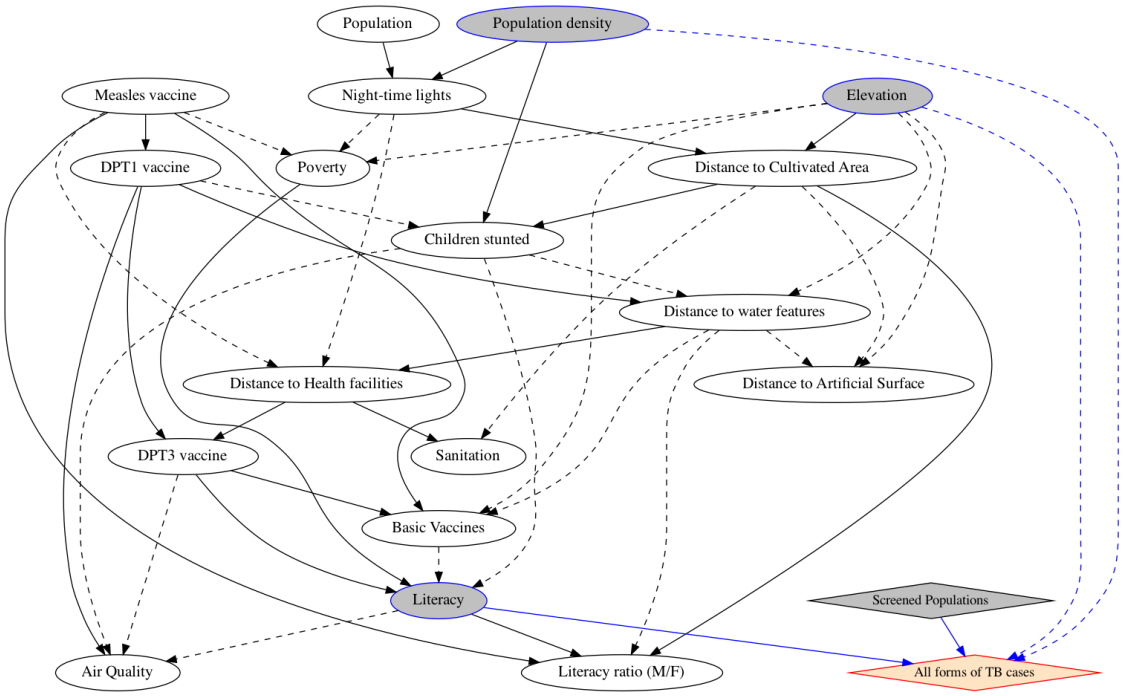

Figure 4: Optimal Additive Bayesian Network Models: for all forms TB cases

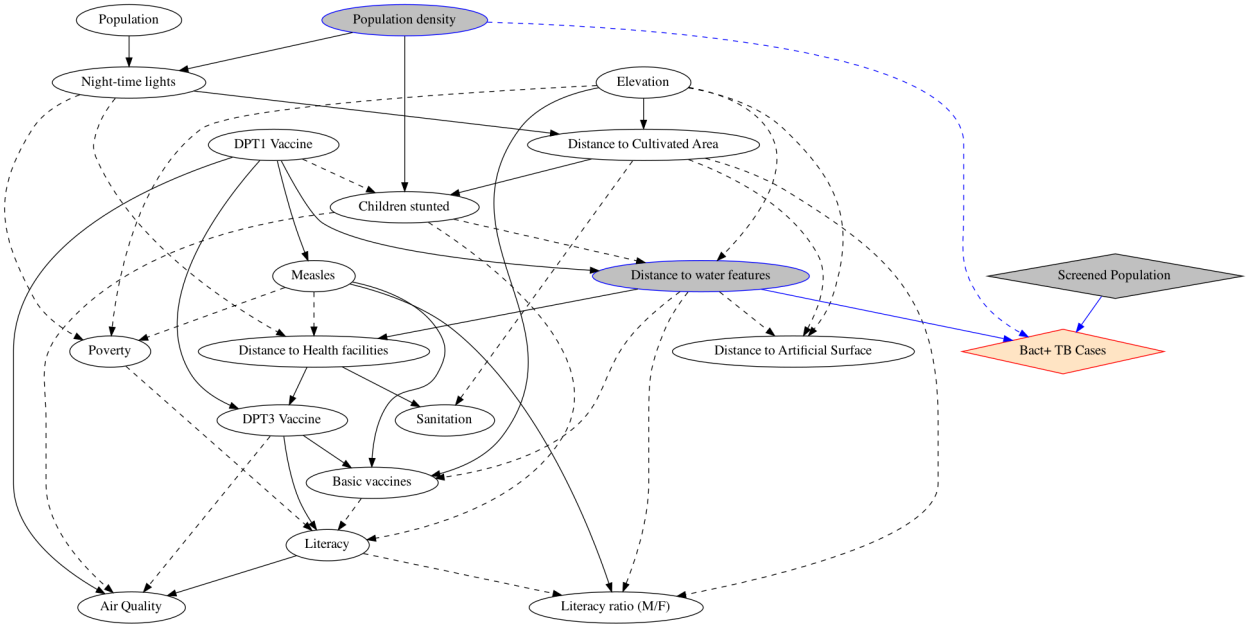

Figure 5: Optimal Additive Bayesian Network Models:: for bacteriological diagnosed TB cases

The area under the parameter has been estimated to check whether the parameter estimates are correct or not. All of the parameters had an area of approximately one under the marginal density (see Figure 8 and 9). Hence, the parameters are well estimated. The parameter estimate with corresponding credible intervals from the optimal DAG is presented in Table 11 and 11. The parameter estimate related to the outcome of interest is presented in Table 5 for both TB cases seniors. The estimates in Table 5 are interpreted as log rate ratio (rate ratio while exponented).

Table 5: Estimates of posterior marginal density: median (log rate ratio) and 95% Credible Interval (CI) ABN Model 1 and ABN Model 2

Parent	**All forms of TB cases**		**Bact+ TB cases**	
(Predictors)	**Estimates**	**95% CI**	**Estimate**	**95% CI**
Intercept	0.6383	(0.5394, 0.732)	-0.2953	(-0.4574, -0.1443)
Population density	-0.341	(-0.4225, -0.2624)	-0.3514	(-0.4926, -0.224)
Literacy	0.3306	(0.255, 0.4034)		
Elevation	-0.1992	(-0.2955, -0.1127)		
Distance to water features			0.2791	(0.1197, 0.4592)
Screened Population	0.0054	(0.0048,0.006)	0.0061	(0.0052, 0.0071)

Based on Figure 4 and Table 5, the overall literacy rate, population density and elevation are directly associated with the all forms of TB cases. The overall literacy rate is positively associated with all tuberculosis rates. A one percent increase of literacy rate at theissen level is associated with 39% (*exp(0.331)=1.39*) increases in the number of tuberculosis diagnosed clinically. Population density is negatively associated with both all forms and bacteriological diagnosed tuberculosis. The increase in the number of people living in a one-kilometer square area by one is associated with a 29 percent (*exp(-0.341)=0.81*) decrease in the rate of all forms of tuberculosis and the rate of bacteriological diagnosed tuberculosis by 30% (*exp(-0.3514)=0.70*). The effect of population density remains the same in both seniors.

Furthermore, elevation is negatively associated with the number of tuberculosis diagnosed either clinically or bacteriologically, while the distance to water features has a positive association with the number of bacteriologically diagnosed tuberculosis. A one-meter increase above the sea level

decreases the rate of all forms of tuberculosis by 18% (*exp(-0.2)=0.82*).

3.4 Prediction

In order to see how well the models are predicting the observations, root mean squared error (RMSE) was computed and the result is presented in Table 6. The predicted rates in the ABN model were obtained by dividing the predicted count by the corresponding observed screened population. However, the ABN model predicted a rate higher than one for seven thiessesns, which is not practically feasible, due to the fact that the model had not the total screened population as an offset variable. These rates were all set to one. As shown in Figure 6, the NB model predicates rate close to the observed compared to the ABN model. The NB has a lower error score compared to the ABN model.

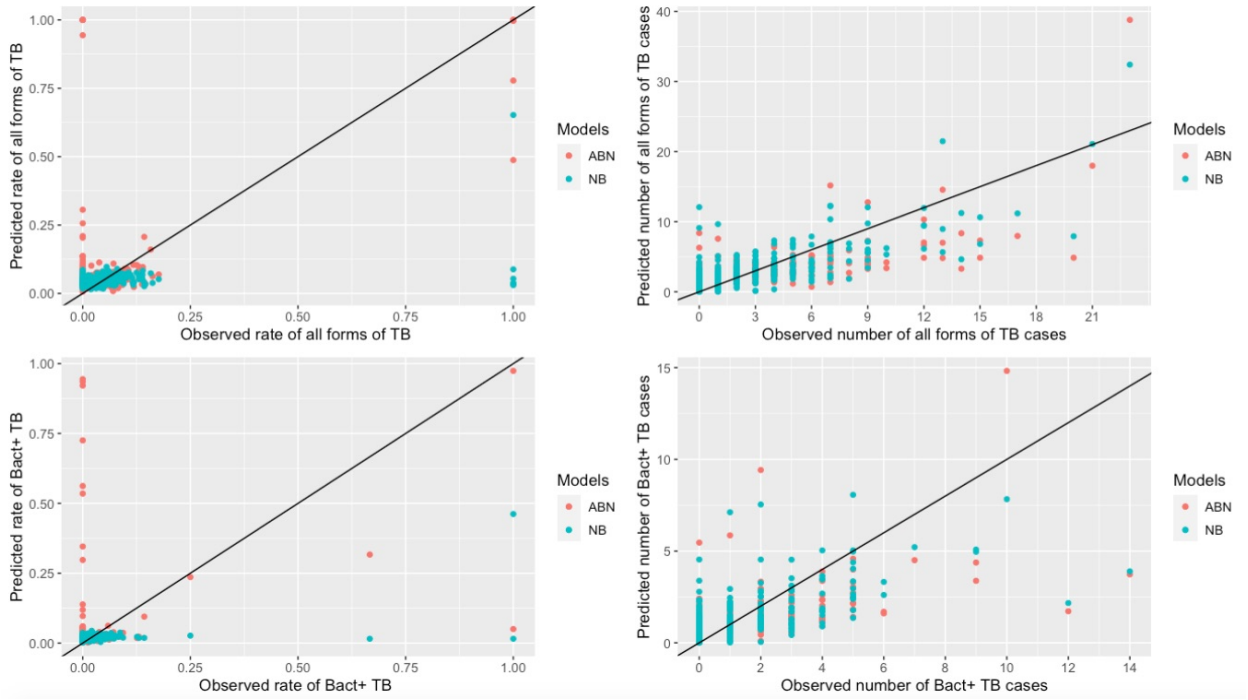

Figure 6: Comparison between Negative Binomial model (NB) and Additive Bayesian network (ABN) for predicted rate (left side) and count (right side) for the all forms of TB (top) and for bacteriological diagnosed TB (bottom)

Table 6: Root mean squared error (RMSE) for the negative binomial and ABN model for each response variable

RSME	NB		ABN	
	Count	Rate	Count	Rate
All forms of TB (total_af)	2.5904	0.1295	2.9197	0.1521
Bact+ TB (total_bpos)	1.5458	0.0791	1.6684	0.1314

4 Discussion

This study was conducted to identify which factors are associated with tuberculosis positivity rate in Pakistan using regression-based approach and Bayesian network approach. It also aimed to compare the results from the regression-based approach with the result from the Bayesian network approach. The effect of environmental, socio-demographic, and access to healthcare-related factors on rate of tuberculosis positivity was investigated in this study. Various analyses were performed to address the study's objective. Several regression-based approaches were used to analyze the data, including a negative binomial model to account for the overdispersion in the Poisson regression and GLMMs for accounting for the clustering nature of the data. It was shown that study data were overdispersed, making the negative binomial model more suitable than the Poisson model. Additionally, the inclusion of the clustering nature of the data did not improve the model, so that standard models (GLMs) were chosen.

Moreover, the data were analyzed using an Additive Bayesian network. The Additive Bayesian Network provides a better understanding of which factors are directly and indirectly associated with TB cases. The factors strongly associated with the response variable have a direct relationship with response in the DAG , as shown in Figure 4 and 5.

The GLM and ABN are more likely to find the same covariate when the associations are strong and highly significant [19, 20]. As mentioned in section 2.2.1 and 2.4, the response variable was different for the regression-based approach (GLM) and the Bayesian network approach (ABN). As a result, it is difficult to make a direct comparison of the covariates estimates in both methods. However, the comparison can be done in terms of the significant association of the covariates with the response. The result from both approaches is presented in Table 7. The overall literacy rate was found to be positively associated with the number of clinically diagnosed tuberculosis cases and indirectly linked to bacteriological TB cases. It was also associated with both rates of tuberculosis positivity. In addition, DPT1 vaccine coverage was indirectly associated with both TB cases.

Table 7: Parameter estimate, standard error (Std.Error) , p-value for GLM, and estimates of the marginal posterior densities (median and 95% credible interval (CI)) with effect direction for ABN model

	All forms of TB cases					Bact+ diagnosed TB cases				
	Negative Binomial		ABN			Negative Binomial		ABN		
	Estimate (Std.Error)	p-value	Effects		Estimate (95% CI)	Estimate (Std.Error)	p-value	Effects		Estimate (95% CI)
			Direct	Indirect				Direct	Indirect	
Intercept	-3.1211 (0.0467)	< 0.0001***			0.638 (0.539,0.732)	-4.0086 (0.0650)	< 0.0001***			-0.295 (-0.457,-0.144)
demog_dpt1	0.3315 (0.0692)	< 0.0001***		X		0.64151 (0.1439)	< 0.0001***		X	
demog_literacy	0.2683 (0.0726)	0.0002***	X		0.331 (0.255, 0.403)	0.2189 (0.0760)	0.0039**		X	
infra_water	-0.4464 (0.0980)	< 0.0001***				-0.2205 (0.1245)	0.0765.	X		0.279 (0.12,0.459)
env_elevation	-0.3690 (0.0884)	< 0.0001***	X		-0.199 (-0.296, -0.113)	-0.3752 (0.1251)	0.0027**		X	
demog_literacy_ inequality	-0.1266 (0.0658)	0.0544.								
demog_ poverty	0.1448 (0.0622)	0.0199*								
infra_ cluster2hcf	0.0837 (0.0516)	0.1051		X						
demog_measles						-0.35240 (0.1131)	0.0018**		X	
demog_childst						0.2250 (0.0761)	0.0031**		X	
env_dst190						0.1271 (0.0603)	0.0350*			
demg_popdens			X		-0.341 (-0.423, -0.262)			X		-0.351 (-0.493,-0.224)
total_scr	offset variable		X		0.005 (0.005,0.006)	offset variable		X		0.006 (0.005,0.007)

Bayesian networks are usually expressed in terms of conditional independence relationship and probabilistic properties, leading the arcs to represent the cause-and-effect relationship [23, 26]. However, the direction of the arcs does not imply a causal-and-effect direction in observation setting. Hence, the result of the ABN model in this study only implies the association among the considered covariates, as the data came from observational study, and the model structure (DAG) was not pre-specified but learned from the dataset.

TB notification rate tends to be higher in areas with better access to healthcare [30]. Hence, the rate of TB positivity tends to be higher as well. This study found that higher coverage percentage of DPT1 and measles vaccines were associated with a higher rate of tuberculosis diagnosed by bacteriological method (see Table 7). More TB patients would be identified by bacteriological method (90%) than using clinical method (40%) in thiessen with high coverage of DPT 1 vaccine. It attributes to the fact that vaccination coverage indicates better access to health care facility and improved uptake of health services. Therefore, the rate of TB diagnosed by bacteriological methods could be higher in the area with better access to healthcare.

Several studies showed that the prevalence of tuberculosis is associated with low-socioeconomic

status [31, 32, 33]. Proportion of stunning can be also considered as indicator of poverty [34, 35]. This study also found that high proportions of people living in poverty and high proportion of stunting were associated with higher TB positivity rate. As a result, TB positivity rates would be higher in areas with a higher proportion of poor people. It could be related to the fact that the high prevalence of TB cases among poor people who are exposed to various diseases. Consequently, there is a high rate of tuberculosis positivity among those screened.

The study found that the overall literacy rate effect remained constant and was associated with tuberculosis positivity. According to the findings, higher rate of tuberculosis positivity was associated with higher literacy rate in thiessens. This could be attributed to literate people being more likely to be concerned about their health and get a medication soon when they feel uncomfortable [36]. On the other hand, they are aware of the prevention mechanism for TB diseases.

The study found that elevation (altitude) was negatively associated with TB positivity rate. The study finding is consistent with the results of the study [37, 38], which found that tuberculosis infection is less common at high altitudes than at low altitudes. Therefore, fewer patients would be identified among those screened, leading to low TB positivity rate. Lower TB case detection rate, on the other hand, could be attributed to limited resources available.

The predicted values in this study are obtained based on a model that was trained on the complete datasets and was intended to find associations. The predicted values might not be good, as shown in Figure 6. The range of possible values for the input variables are wider than those of the original input variables; the prediction could not be good since the model has not been seen for those values. Because of this extrapolation, the prediction for the other thiessesns was not performed. The prediction could be improved by using a predictive model. Therefore, it is essential to use a model that focuses on prediction rather than statistical inference (i.e., finding an association) to predict future outcomes accurately [39]. The data would be separated into training and testing dataset to develop a predictive model. The model building and its performance evaluation would be performed separately. A prediction error matric thus can be computed from the testing dataset.

The prediction in this study can be improved by suing a predictive model such as support vector machine (SVM) or random forest (RF) methods. The SVM requires a transformation of the rate response. The response rate could be transformed into a logit scale containing any real number. After predicting the *logit* scale value, the prediction for the rate would done using back-transformation (i.e., *expit*). A RF model can be used directly without data transformation. However, these models also suffer from extrapolation issues [40].

This study has the following limitations: First, the results from the regression-based and Bayesian network approaches are not directly comparable because the response variable differed. Second, due to the computation time constraint, the robustness of the ABN model was not checked; thus, the result should be carefully interpreted under this condition.

5 Conclusion and Further work

This study examined the effect of various socio-demographic, environmental, and access to healthcare-related factors on the rate of TB positivity. It allows us to distinguish which factors are associated with tuberculosis positivity rate. TB positivity is associated with overall literacy rate, literacy inequality between males and females, poverty, stunting, elevation, distance to water features, and vaccine coverage. The additive Bayesian network also allows us to identify factors directly and indirectly associated with TB cases. TB burden is directly associated with population density, literacy rate, elevation and distance to water features, while it associated indirectly with poverty, vaccine coverage, and near access of health facility. Expanding access to health care facilities would aid in the identification of Tuberculosis patients. The study will help the society in having awareness towards the effect of socio-demographic, environmental and healthcare-related factors associated with the rate of diagnosing TB cases.

It is recommended to use predictive models for estimating the TB positivity rate across all thiessens. It also suggests conducting additional spatial analyses to detect high TB risk areas, investigate which factors are associated with TB positivity rates, and compare the results with the regression approach.

6 Possible Drawback of the used Method

The ABN has some limitations. The first limitation is computational intensive for performing bootstrap procedure to minimize overfitting. The second limitation of the ABN method is that it limits the types of variables allowed in the model. Only binary, Poisson, and normally distributed continuous variables are allowed in the model [19, 20]. All continuous variables (including these expressed in percentage) were transformed in standardised scale (using the default function *fitAbn()* [19]) to approximate normal distribution. However, this workaround makes the interpretation of the posterior marginal estimate in the transformed scale harder. Furthermore, with overdispersed data (as shown in the GLM part), there was no option for choosing an appropriate distribution for the response variable (count variable), such as a negative binomial distribution rather than a Poisson distribution.

Another drawback of the ABN approach does not incorporate an offset variable. The inclusion of offset variables in the ABN model is currently not feasible in this study. Since an offset variable has a fixed coefficient of one in the model, there is no way to shrink the coefficient of the offset variable to one in ABN. It makes direct comparisons of covariate estimates from the GLM and ABN models harder.

Ethical Consideration

The screening data for this study obtained from the EPCON company in Antwerp, Belgium. The data were aggregated at the thiessen level, making it impossible to have information about individuals or families. As a result, there was no ethical concern regarding these subjects.

Stakeholder Awareness

The study was conducted with the collaboration of the EPCON company. One of the key stockholders in the study is the EPCON company, which aimed to understand TB epidemiology and find a way to improve the existing model. The study would help the NTP in increasing the testing rate.

References

[1] WHO. Global Tuberculosis Report 2020. https://www.who.int/news-room/fact-sheets/detail/tuberculosis, 2021. accessed: 24.02.2022.

[2] CDC. Basic TB Facts. https://www.cdc.gov/tb/topic/basics/default.htm, 2022. accessed: 23.05.2022.

[3] WHO. *Global tuberculosis report 2020*. World Health Organization, 2021. Geneva.

[4] WHO. Implementing the End TB Strategy. https://www.who.int/westernpacific/activities/implementing-the-end-tb-strategy, 2022. accessed: 13.05.2022.

[5] Lonnroth KU, Uplekar M, and Ottmani S. An action framework for higher and earlier TB case detection: Background document for DOTS Expansion Working Group. *STOP TB partnership. Geneva*, pages 13–14, 2009.

[6] Somma D, Thomas BE, Karim F, Kemp J, Arias N, Auer C, Gosoniu GD, Abouihia A, and Weiss MG. Gender and socio-cultural determinants of TB-related stigma in Bangladesh, India, Malawi and Colombia: Special section on gender and TB. *The International Journal of Tuberculosis and Lung Disease*, 12(7):856–866, 2008.

[7] Alba S, Rood E, Mecatti F, Ross JM, Dodd PJ, Chang S, Potgieter M, Bertarelli G, Henry NJ, LeGrand KE, Trouleau W, et al. TB Hackathon: Development and Comparison of Five Models to Predict Subnational Tuberculosis Prevalence in Pakistan. *Tropical medicine and infectious disease*, 7(1):13, 2022.

[8] Ngwenya M. Factors contributing to low Tuberculosis case finding in Zimbabwe: New challenges demand innovative approaches. 2015.

[9] NTP. Tuberculosis profile: Pakistan. https://ntp.gov.pk/tb-profile-pakistan/, 2022. accessed: 15.05.2022.

[10] StopTB Partnership. Tuberculosis situation in 2020:Pakistan, Low-middle income. https://www.stoptb.org/static_pages/PAK_Dashboard.html, 2022. accessed: 18.05.2022.

[11] National Institute of Health. National TB control program of Pakistan. https://www.nih.org.pk/national-tb-control-program/, 2022. accessed: 05.03.2022.

[12] NTP. Annual report 2012. *Ministry of National Health Services, Regulation Coordination,Islamabad, Pakistan*, 2012.

[13] NTP. Annual Report 2016. *Ministry of National Health Services, Regulation Coordination, Islamabad, Pakistan*, 2016.

[14] NTP. Annual Report: 2019. *Ministry of National Health Services, Regulation Coordination,Islamabad, Pakistan*, 2019.

[15] WHO. Chest Radiography In Tuberculosis Detection: Summary of current WHO recommendations and guidance on programmatic approaches. 2016. Geneva.

[16] Agresti A. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, 2007.

[17] Hilbe JM. *Modeling Count Data*. Cambridge University Press, 2014.

[18] Molenberghs G. and Verbeke G. *Models for Discrete Longitudinal Data*. Springer Inc., New York, 2005.

[19] Kratzer G, Lewis FI, Comin A, Pittavino M, and Furrer R. Additive Bayesian network modelling with the R package Abn. *arXiv preprint arXiv:1911.09006*, 1, 2016.

[20] Pittavino M. *Additive Bayesian networks for multivariate data: parameter learning, model fitting and applications in veterinary epidemiology*. PhD thesis, University of Zurich, 2016.

[21] Faraway JJ. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, 2016.

[22] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[23] Pearl J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[24] Arora P, Boyne D, Slater JJ, Gupta A, Brenner DR, and Druzdzel MJ. Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Value in Health*, 22(4):439–445, 2019.

[25] Koller D and Friedman N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[26] Scutari M and Denis JB. *Bayesian networks: with examples in R*. Chapman and Hall/CRC, 2021.

[27] Hornỳ M. Bayesian networks. *Boston University School of Public Health*, 17, 2014.

[28] Adnan Darwiche. *Modeling and reasoning with Bayesian networks*. Cambridge university press, 2009.

[29] Jensen FV and Nielsen TD. *Bayesian networks and decision graphs*, volume 2. Springer, 2007.

[30] Dangisso MH, Datiko DG, and Lindtjørn B. Accessibility to tuberculosis control services and tuberculosis programme performance in southern Ethiopia. *Global health action*, 8(1):29443, 2015.

[31] Jiamsakul A, Lee MP, Nguyen KV, Merati TP, Cuong DD, Ditangco R, Yunihastuti E, Ponnampalavanar S, Kiertiburanakul F, Zhang S, and Avihingasanon A. Socio-economic status and risk of tuberculosis: a case-control study of HIV-infected patients in Asia. *The International Journal of Tuberculosis and Lung Disease*, 22(2):179–186, 2018.

[32] Wu J and Dalal K. Tuberculosis in Asia and the pacific: the role of socioeconomic status and health system development. *International journal of preventive medicine*, 3(1):8, 2012.

[33] Dheeraj Gupta, Kshaunish Das, T Balamughesh, N Aggarwal, and Surinder K Jindal. Role of socio-economic factors in tuberculosis prevalence. *Indian Journal of Tuberculosis*, 51(1):27–32, 2004.

[34] Gross R, Schultink W, and Sastroamidjojo S. Stunting as an indicator for health and wealth: an Indonesian application. *Nutrition research*, 16(11-12):1829–1837, 1996.

[35] Setboonsarng S. Child malnutrition as a poverty indicator: an evaluation in the context of different development interventions in Indonesia. Technical report, ADBI Discussion Paper, 2005.

[36] Andrus MR and Roth MT. Health Literacy: A Review. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 22(3):282–302, 2002.

[37] Gelaw YA, Yu W, Magalhães RJ, Assefa Y, and Williams G. Effect of temperature and altitude difference on tuberculosis notification: a systematic review. *Journal of global infectious diseases*, 11(2):63, 2019.

[38] Mansoer JR, Kibuga DK, and Borgdorff MW. Altitude: a determinant for tuberculosis in Kenya? *The International Journal of Tuberculosis and Lung Disease*, 3(2):156–161, 1999.

[39] James G, Witten D, Hastie T, and Tibshirani R. *An introduction to statistical learning*, volume 112. Springer, 2013.

[40] Saeed F, Al-Hadhrami T, Mohammed E, and Mohammed E. *Advances on Smart and Soft Computing: Proceedings of ICACIn 2020*. Springer, 2021.

Appendices

Appendix A: Tables

Table 8: Variable Descriptions

Categorizations	Codes	Descriptions
Access to Health Care	infra_cluster2hcf	distance from a thiessen cluster to the nearest healthcare facility
	demog_basicvaccs	average percentage of basic vaccinations
	demog_dpt1	average percentage of DPT 1 vaccinations
	demog_dpt3	average percentage of DPT 3 vaccinations
	demog_measles	average percentage of measles vaccination
Socio-Demographic	demog_poverty	proportion of people living in poverty
	demog_literacy	proportion of literate people
	demog_literacy_inequality	male to female literacy ratio
	infra_dmsp	DMSP-OLS night-time lights in Pakistan for the year 2000
	pop	population
	demog_popdens	population density of a given thiessen
	infra_water	average distance to ESA-CCI-LC water feature (kilometers)
	infra_sanitation	average percentage of open defecation
	demog_childst	average percentage of children stunting
Environmental factors	env_elevation	elevation
	env_dst011	distance to ESA-CCI-LC cultivated area edges 2015
	env_dst190	distance to ESA-CCI-LC artificial surface edges 2015
	env_aqhchoq1	average in-door house air quality for first quarter (January, February, March)
	env_aqhchoq2	average in-door house air quality for the second quarter (April, May, June)
	env_aqhchoq3	average in-door house air quality for the third quarter (July, August, September)
	env_aqhchoq4	average in-door house air quality for the fourth quarter (October, November, December)
	env_aqhchoq	average in-door house air quality for a year
Response Variable	total_af	total all forms of TB cases
	total_bpos	total bacteriological diagnosed TB cases
	total_scr	total screened population

Table 9: Parameter estimate, standard error (Std.Error), and p-value for all forms of TB positivity rate: Poission and GLMM: Poisson

	Poisson			**GLMM: Poisson**		
	Estimate	**Std.Error**	**P-value**	**Estimate**	**Std.Error**	**p-value**
(Intercept)	-3.1376	0.0340	< 0.0001***	-3.1342	0.0549	< 0.0001***
env_elevation	-0.3013	0.0749	< 0.0001***	-0.3347	0.0878	< 0.0001***
infra_cluster2hcf	0.0602	0.0375	0.1082	0.2783	0.0703	0.0001
demog_dpt1	0.2637	0.0576	< 0.0001***	0.0953	0.0426	0.0253*
infra_water	-0.3571	0.0870	< 0.0001***	-0.3999	0.0928	< 0.0001***
demog_literacy	0.2633	0.0528	< 0.0001***	0.2047	0.0681	0.0026**
demog_literacy_inequality	-0.1306	0.0480	0.0065***	-0.0853	0.0555	0.1246
demog_popdens	-0.0757	0.0370	0.0407*	-0.0633	0.0493	0.1985
demog_poverty	0.1274	0.0489	0.0091**	0.0259	0.0722	0.7196
			σ^2	0.0628	LRT=14.06246	< 0.0001@

$***pvalue < 0.001$; $**p - value < 0.01$; $*p - 0.05$; $.pvalue < 0.1$;

@*pvalue* based on likelihood ratio test

Table 10: Parameter estimate, standard error (Std.Error), and p-value for all forms of TB positivity rate: GLMM: Negative Binomial

	Estimate	**Std.Error**	**p-value**
Intercept	-3.1263	0.0513	< 0.0001***
env_elevation	-0.3695	0.0895	< 0.0001***
demog_dpt1	0.3268	0.0715	< 0.0001***
infra_cluster2hcf	0.0922	0.0548	0.0924
infra_water	-0.4460	0.0972	< 0.0001***
demog_literacy	0.2599	0.0775	< 0.0008***
demog_literacy_inequality	-0.1210	0.0677	0.0741
demog_poverty	0.1246	0.0711	0.0797
θ	0.0722	0.7196	
σ_u^2	0.01402	LRT=0.5825	0.2226 @

$***pvalue < 0.001$; $**p - value < 0.01$; $*p - 0.05$; $.pvalue < 0.1$;

@*pvalue* based on likelihood ratio test

Table 11: Parameter estimate, 95% credible interval from the Global Optimal DAG of ABN Model 1

Child	Parents	Estimates (95% CI)	Child	Parents	Estimates (95% CI)	Child	Parents	Estimates (95% CI)
env_aqhchoq	demog_childst	-0.1462 (-0.2125,-0.0799)	infra_sanitation	env_dst011	-0.2305 (-0.3471,-0.118)	demog_literacy	demog_basicvaccs	-0.561 (-0.6768,-0.4493)
	demog_dpt1	0.7052 (0.6241,0.7862)		infra_cluster2chf	0.2385 (0.1233,0.3536)		demog_childst	-0.3774 (-0.4597,-0.2981)
	demog_dpt3	-0.281 (-0.3604,-0.2045)	infra_water	env_elevation	-0.6346 (-0.7,-0.5715)		demog_dpt3	0.2175 (0.1037,0.3312)
	demog_literacy	0.5381 (0.475,0.6011)		demog_childst	-0.2284 (-0.2965,-0.1626)		demog_poverty	-0.465 (-0.5474,-0.3857)
env_dst011	env_elevation	0.186 (0.0919,0.28)		demog_dpt1	0.3664 (0.2987,0.434)	demog_literacy	env_dst011	0.1778 (0.0935,0.2621)
	infra_dmsp	0.5528 (0.4594,0.6462)	demog_basicvaccs	env_elevation	-0.3471 (-0.441,-0.2566)		infra_water	-0.2121 (-0.3031,-0.1244)
env_dst190	env_dst011	-0.7805 (-0.8463,-0.717)		infra_water	-0.2181 (-0.3253,-0.1147)	inequality_	demog_literacy	-0.6808 (-0.7666,-0.598)
	env_elevation	-0.3059 (-0.3956,-0.2194)		demog_dpt3	0.503 (0.4165,0.5895)		demog_measles	0.2021 (0.1141,0.2901)
	infra_water	-0.3597 (-0.4489,-0.2737)		demog_measles	0.4531 (0.3729,0.5333)	demog_poverty	env_elevation	-0.2957 (-0.3792,-0.2152)
infra_cluster2hcf	infra_dmsp	-0.5231 (-0.6174,-0.4322)	demog childst	env_dst011	0.204 (0.0953,0.3126)		infra_dsmp	-0.5187 (-0.602,-0.4384)
	demog_basicvaccs	0.3179 (0.2211,0.4147)		demog_popdens	-0.4516 (-0.5622,-0.3447)		demog_measles	-0.4805 (-0.5645,-0.3993)
	demog_measles	-0.3889 (-0.4888,-0.2925)		demog_dpt1	-0.3743 (-0.4756,-0.2767)	total_af	env_elevation	-0.1992 (-0.2955,-0.1127)
infra_dsmp	demog_popdens	0.7377 (0.6697,0.8058)	demog_dpt3	infra_cluster2hcf	0.2951 (0.2181,0.3721)		demog_literacy	0.3306 (0.255,0.4034)
	pop	0.2389 (0.1714,0.3065)		demog_dpt1	0.7207 (0.6436,0.7978)		demog_popdens	-0.341 (-0.4225,-0.2624)
			demog_dpt1	demog_meales	0.8065 (0.7384,0.8746)		total_scr	0.0054 (0.0048,0.006)

Table 12: Parameter estimate, standard error (Std.Error), and corresponding p-value for Poisson Model: total_bpos

	Estimate	**Std.Error**	**P-Value**
Intercept	-4.0154	0.0547	$< 0.0001 * **$
env_aqhchoq	-0.2048	0.1048	0.0510.
env_dst190	0.1278	0.0493	0.0117*
env_elevation	-0.2269	0.0934	0.0084**
demog_dpt1	0.7142	0.14	$< 0.0001 * **$
demog_dpt3	-0.1864	0.0937	0.0476*
demog_measles	-0.2768	0.1005)	0.0054**
demog_childst	0.2283	0.0671	$< 0.0001 * *$
demog_literacy	0.2128	0.0919)	0.0198*
demog_literacy_inequality	-0.1053	0.0717	0.1397

$***pvalue < 0.001$; $** p - value < 0.01$; $*p - 0.05$; $.pvalue < 0.1$

Table 13: Parameter estimate from the Global Optimal DAG of ABN Model 2

Child	Parents	Estimates (95% CI)	Child	Parents	Estimates (95% CI)	Child	Parents	Estimates (95% CI)
env_aqhchoq—	demog_childst	-0.1462 (-0.2125,-0.0799)	infra_sanitation	env_dst011	-0.2305 (-0.3471,-0.118)	demog_literacy	demog_basicvaccs	-0.561 (-0.6768,-0.4493)
	demog_dpt1	0.7052 (0.6241,0.7862)		infra_cluster2chf	0.2385 (0.1233,0.3536)		demog_childst	-0.3774 (-0.4597,-0.2981)
	demog_dpt3	-0.281 (-0.3604,-0.2045)	infra_water	env_elevation	-0.6346 (-0.7,-0.5715)		demog_dpt3	0.2175 (0.1037,0.3312)
	demog_literacy	0.5381 (0.475,0.6011)		demog_childst	-0.2284 (-0.2965,-0.1626)		demog_poverty	-0.465 (-0.5474,-0.3857)
env_dst011	env_elevation	0.186 (0.0919,0.28)		demog_dpt1	0.3664 (0.2987,0.434)	demog_literacy	env_dst011	0.1778 (0.0935,0.2621)
	infra_dmsp	0.5528 (0.4594,0.6462)	demog_basicvaccs	env_elevation	-0.3471 (-0.441,-0.2566)		infra_water	-0.2121 (-0.3031,-0.1244)
env_dst190	env_dst011	-0.7805 (-0.8463,-0.717)		infra_water	-0.2181 (-0.3253,-0.1147)	inequality_	demog_literacy	-0.6808 (-0.7666,-0.598)
	env_elevation	-0.3059 (-0.3956,-0.2194)		demog_dpt3	0.503 (0.4165,0.5895)		demog_measles	0.2021 (0.1141,0.2901)
	infra_water	-0.3597 (-0.4489,-0.2737)		demog_measles	0.4531 (0.3729,0.5333)	demog_poverty	env_elevation	-0.2957 (-0.3792,-0.2152)
infra_cluster2hcf	infra_dmsp	-0.5231 (-0.6174,-0.4322)	demog childst	env_dst011	0.204 (0.0953,0.3126)		infra_dsmp	-0.5187 (-0.602,-0.4384)
	infra_water	0.3179 (0.2211,0.4147)		demog_popdens	-0.4516 (-0.5622,-0.3447)		demog_measles	-0.4805 (-0.5645,-0.3993)
	demog_measles	-0.3889 (-0.4888,-0.2925)		demog_dpt1	-0.3743 (-0.4756,-0.2767)	demog_measles	demog_dpt1	0.8065 (0.7384,0.8746)
infra_dsmp	demog_popdens	0.7377 (0.6697,0.8058)	demog_dpt3	infra_cluster2hcf	0.2951 (0.2181,0.3721)	total_bpos	demog_popdens	0.3306 (0.255,0.4034)
	pop	0.2389 (0.1714,0.3065)		demog_dpt1	0.7207 (0.6436,0.7978)		infra_water	-0.341 (-0.4225,-0.2624)
							total_scr	0.0054 (0.0048,0.006)

Appendix B: Figures

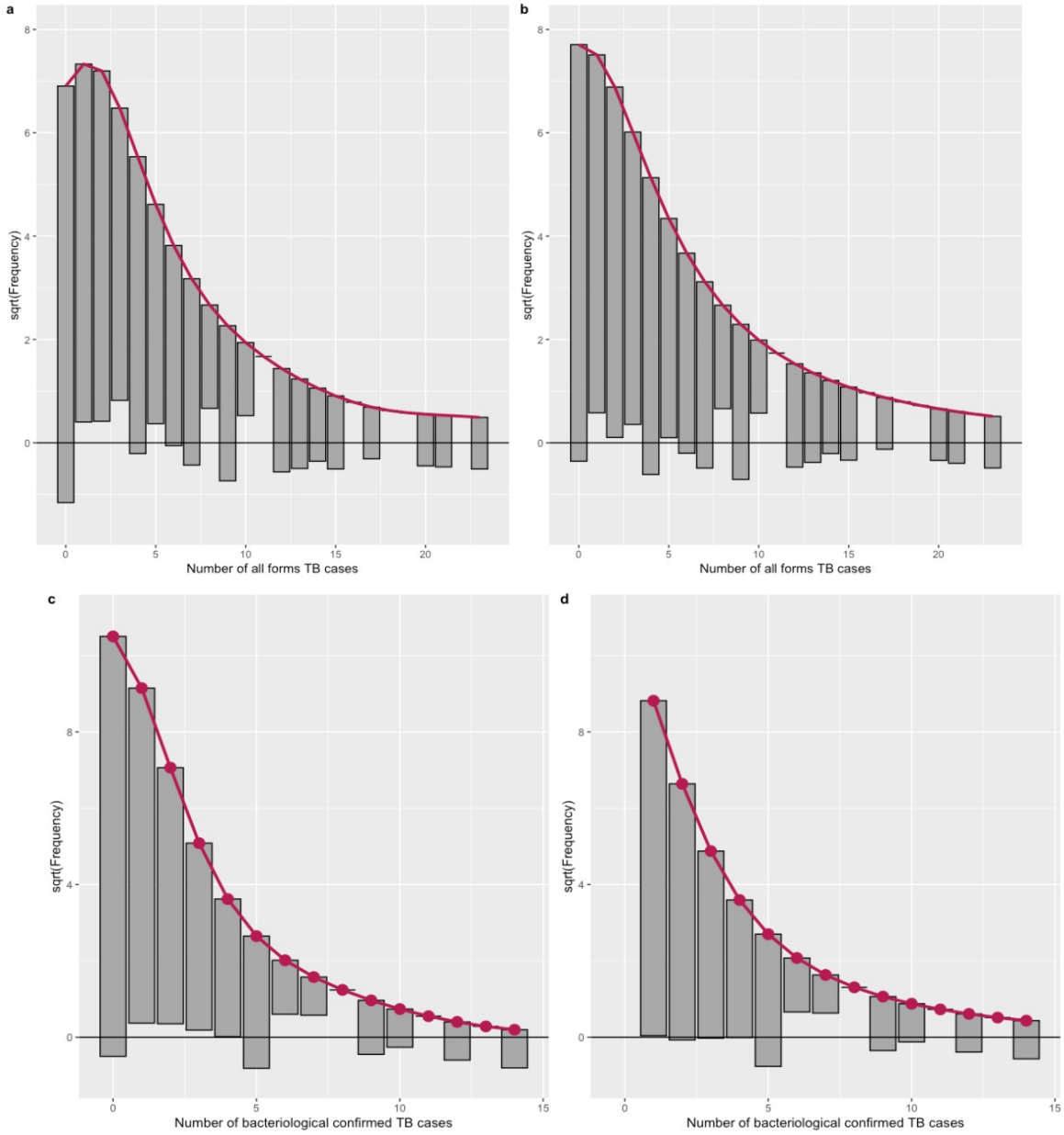

Figure 7: Rootogram graph for all forms of TB positivity rate (top) and Bact+ TB positivity rate (bottom) for Poisson and Negative Binomial model (left to right)

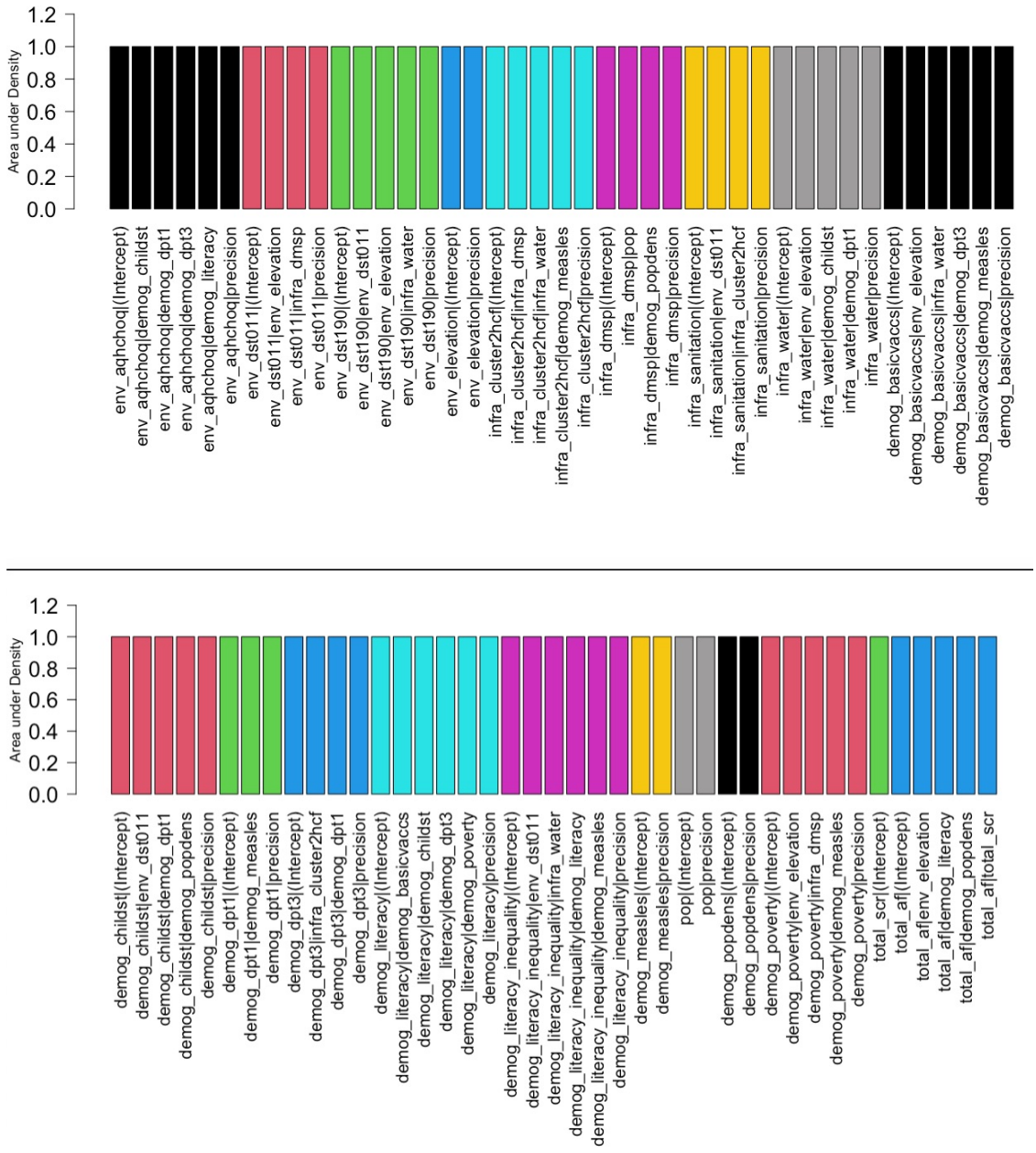

Figure 8: Approximated area under the marginal densities of parameters: ABN Mode 1

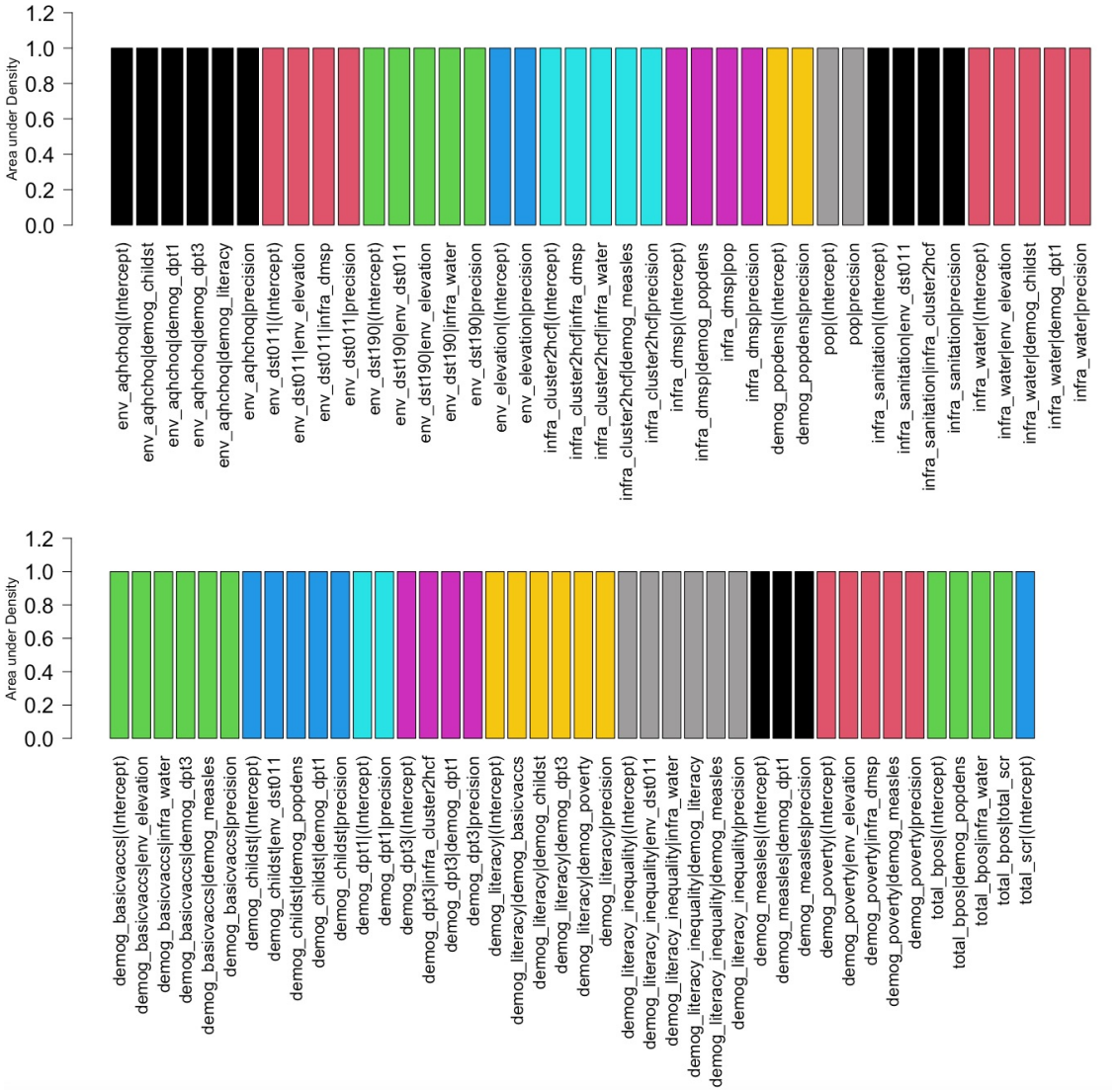

Figure 9: Approximated area under the marginal densities of parameters: ABN Mode 2

Appendix C: R code

```
######################## Regression approach (GLM) ################################
# to install packages
ipak<- function(pkg){new.pkg<- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg)) install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)}
packages<- c("ggplot2","dplyr", "tidyverse", "ggplot2","patchwork","kableExtra",
             "tmap","ggcorrplot","stringr","patchwork","vcd","readr","hablar",
             "AER","countreg","cowplot","emdbook","lmer"); ipak(packages)
# load dataset
Data_new<-read.csv(file.choose())
################ all forms of TB cases ##########################
Data_new%>%summarise(Mean = round(mean(total_af),2),Variance =round(var(total_af),2),
Std.Dev=round(sd(total_af),2), ratio = round(var(total_af) / mean(total_af),2))
# Distribution of total_af using barplot
G2<-Data_new%>%group_by(total_af)%>%dplyr::summarise(N=n())%>%
               mutate(Percentage=N/sum(N))
P1<-ggplot(G2, aes(x=as.factor(total_af), y=N)) + geom_bar(stat = "identity",
fill = "lightblue")+labs(y = "Frequency", x = "Number of all forms TB cases")+
  scale_y_continuous( breaks = c(seq(0,70, by=5)))+theme(text = element_text(size = 14))
model.taf<-glm(total_af~1,offset=log(total_scr),family = poisson,data=Data_new)
model.nb.taf<-glm.nb(total_af~1+offset(log(total_scr)),link=log,data=Data_new)
root.pois.taf.1<-countreg::rootogram(model.taf, max = max(Data_new$total_af)
                      ,main = " Poisson", xlab="Number of all forms TB cases")
root.nb.taf.1<-countreg::rootogram(model.nb.taf, max = max(Data_new$total_af)
              ,main = " Negative Binomial",xlab="Number of all forms TB cases")
P2<-plot_grid(autoplot(root.pois.taf.1) + ylim(-2,9),autoplot(root.nb.taf.1) +
               ylim(-2,9), ncol = 2,labels="auto",label_size = 12)
P1+P2
######### Poisson model ##################
```

```
model.taf.p1<-glm(total_af~env_elevation + infra_cluster2hcf + demog_dpt1+
     infra_water + demog_literacy +demog_literacy_inequality +
     demog_popdens + demog_poverty, family = poisson, offset = log(total_scr),data = Data_new)
AIC(model.taf.p1);BIC(model.taf.p1);(LogL.taf.p1<-(-2*logLik(model.taf.p1)))
(summary(model.taf.p1))
#overdispersion parameters
(sum(residuals(model.taf.p1, type = "pearson")^2) / model.taf.p1$df.residual)
# comparison between Poisson with GLMM:Poisson
model.taf.pglmm<-glmer(total_af~env_elevation + infra_cluster2hcf + demog_dpt1+
    infra_water + demog_literacy +demog_literacy_inequality + demog_popdens +
    demog_poverty+(1|tehsil_id), family = poisson, offset = log(total_scr),data = Data_new)
Lrt_p_GLMM<-(-2*logLik(model.taf.p1)-(-2*logLik(model.taf.pglmm)))
pchibarsq(Lrt_p_GLMM,df=1,mix=0.5,lower.tail = F)
########## Negative Binomial ####################
model.taf.nb1<-glm.nb(total_af~env_elevation + infra_cluster2hcf + infra_water+ demog_dpt1 +
  demog_literacy+ demog_literacy_inequality+ demog_poverty + offset(log(total_scr)),
  data=Data_new,link=log)
AIC(model.taf.nb1);BIC(model.taf.nb1);(LogL.taf.nb1<-(-2*logLik(model.taf.nb1)))
summary(model.taf.nb1)
### test for overdispersion parameter
model.taf.nb.p<-glm(total_af~env_elevation + infra_cluster2hcf + infra_water + demog_dpt1 +
  demog_literacy+ demog_literacy_inequality + demog_poverty + offset(log(total_scr)),
  data=Data_new,family=poisson(link="log"))
Lrt_NB_p<-(-2*logLik(model.taf.nb.p)-LogL.taf.nb1)
pchibarsq(Lrt_NB_p,df=1,mix=0.5,lower.tail = F)
# comparison between negative binomial with GLMM:NB
model.taf.gNB<-glmer.nb(total_af~env_elevation +demog_dpt1 +infra_cluster2hcf+ infra_water +
  demog_literacy+ demog_literacy_inequality +demog_poverty+
  (1|tehsil_id), data = Data_new, offset = log(total_scr),control=glmerControl(optimizer="boby
AIC(model.taf.gNB);BIC(model.taf.gNB);summary(model.taf.gNB)
```

```
(Lrt_NB.GLMM<-LogL.taf.nb1-(-2*logLik(model.taf.gNB)))
(pchibarsq(Lrt_NB.GLMM,df=1,mix = 0.5,lower.tail = F))
## GLMM:NB with GLMM:Poisson
model.taf.gNB.p<-glmer(total_af~env_elevation + demog_dpt1 +infra_cluster2hcf+ infra_water +
  demog_literacy + demog_literacy_inequality +demog_poverty+(1|tehsil_id), data = Data_new,
  family=poisson,offset=log(total_scr),control=glmerControl(optimizer="bobyqa"))
(Lrt_GLMMNB_GLMMP<-(-2*logLik(model.taf.gNB.p))-(-2*logLik(model.taf.gNB)))
(pchibarsq(Lrt_GLMMNB_GLMMP,df=1,mix = 0.5,lower.tail = F))
## rootgoram graph poisson and negative binomial
root.pois.taf<-countreg::rootogram(model.taf.p1, max = max(Data_new$total_bpos)
                      ,main = " Poisson",xlab="Number of all forms of TB cases")
root.nb.taf<-countreg::rootogram(model.taf.nb1, max = max(Data_new$total_bpos)
                    ,main = " Poisson",xlab="Number of all forms of TB cases")
(P3<-plot_grid(autoplot(root.pois.taf) + ylim(-2,9),autoplot(root.nb.taf) +
               ylim(-2,9), ncol = 2,labels="auto",label_size = 12))
##############################################################################
###################   Additive Bayesian Network Model  ########################
##############################################################################
# package install
packages<- c("readxl","lubridate","abn","mice","dplyr","coda", "ggpubr", "tidyverse",
    "ggplot2","patchwork","kableExtra","tmap","stringr","patchwork","vcd","readr","hablar",
    "AER","cowplot","reshape2","abind","grid","rjags","Cairo","DiagrammeR");ipak(packages)
# Data load
data<-read.csv(file.choose())
names(data); data<-data[,-1];summary(data)
dists<- list(env_aqhchoq = "gaussian",env_dst011 = "gaussian",env_dst190 = "gaussian",
             env_elevation = "gaussian",infra_cluster2hcf= "gaussian",infra_dmsp= "gaussian",
             demog_popdens= "gaussian",pop= "gaussian",infra_sanitation= "gaussian",
             infra_water= "gaussian",demog_basicvaccs= "gaussian",demog_childst= "gaussian",
             demog_dpt1= "gaussian", demog_dpt3= "gaussian",demog_literacy= "gaussian",
```

```
              demog_literacy_inequality= "gaussian", demog_measles= "gaussian",
              demog_poverty="gaussian", total_af="poisson",total_scr="poisson")
dists<- dists[names(data)]
#  loop for discovering needed network complexity or to find globally best DAG
mycache<-dag<-fabn<-list();mlik.values<-c(1:7)
for (i in 1:8){
  max.par<-i
  mycache[[i]]<-buildscorecache(data.df=data, data.dists=dists,method="bayes",
                           dag.banned = ~total_scr|., max.parents = max.par)
  dag[[i]]<-mostprobable(score.cache = mycache[[i]],score = "mlik")
  fabn[[i]]<-fitabn(object = dag[[i]],method = "bayes"); mlik.values[i]<-fabn[[i]]$mlik}
mlik.values<-c(-15198,-14950.45,-14855.76,-14846,-14845.76,-14845.76,-14845.76)
XX<-data.frame(cbind(mlik.values,max.par=c(1:7)))
jpeg("Network Score_taf1.jpeg", quality = 95)
ggplot(XX,aes(x=max.par,y=mlik.values))+geom_point()+geom_line()+
  scale_y_continuous(breaks = seq(-15198,-14845.76,50))+
  geom_vline(xintercept = which(mlik.values==max(mlik.values))[1],col="blue")+
  scale_x_continuous(breaks = seq(1,7,1))+
  ylab("Network score: marginal likelihood")+xlab("Number of parent")+
  theme(text = element_text(size=16,color = "black"),axis.text.x = element_text( hjust=1))
dev.off()
#plot the DAG with Graphviz
mycache<-mycache[[5]];data<- data; dists<-dists; max.par<- 5
dag<-dag[[5]]
toGraphviz(dag = dag$dag, data.df = data, data.dists = dists,
           outfile = paste0("OptimalDAG",max.par,".dot"))
system("dot -Tpng -o OptimalDAG5.png OptimalDAG5.dot")
#########################################################################
marginals estimate for each and every parameter in the model
#########################################################################
```

```
marg.par<-fitabn(dag=dag$dag,data.df=data,data.dists=dists,compute.fixed=TRUE,
                 n.grid=1000)
# Extract values marginal values
marg.new<- marg.par$marginals[[1]]
for(i in 2:length(marg.par$marginals)){marg.new<- c(marg.new, marg.par$marginals[[i]])}
### Visually inspect the marginal posterior distributions of the parameters
CairoPDF("SummaryOfOptimalExactDAG_taf.pdf");
for(i in 1:length(marg.par$marginals)){
    current.node<- marg.par$marginals[i]; child.node.name<- names(marg.par$marginals)[i]
    current.node<- current.node[[1]]
    for(j in 1:length(current.node) ) {
        parent.node.name<-names(current.node)[j]; current.param<- current.node[[j]]
    plot(current.param,type="l",main=paste(child.node.name, ":", parent.node.name), cex=0.7)}}
dev.off()
#to calculate the area under the density curve
marg.density<- marg.par$marginals[[1]]
for (i in 2:length(marg.par$marginals)){marg.density<- c(marg.density, marg.par$marginals[[i]]
AUC<- rep(NA, length(marg.density)) ; names(AUC)<- names(marg.density)
for(i in 1:length(marg.density)) {
  tmp<- spline(marg.density[[i]]); AUC[i]<- sum(diff(tmp$x[-length(tmp$x)])*tmp$y[-1])}
# color for the child (response variable)
colors_auc<-NA
for(i in 1:length(marg.par$marginals)){
  current.node<- marg.par$marginals[i]
  child.node.name<- names(marg.par$marginals)[i]; current.node<- current.node[[1]]
  for(j in 1:length(current.node) ) {
    auc_col<-rep(i,length(current.node))}; colors_auc<-c(colors_auc,auc_col)}
colors_auc<-colors_auc[-1]
par(mfrow=c(1,1));par(las=2, mar=c(21,4,2.0,0.5));rr<-data.frame(AUC)
rr<-data.frame(AUC,parameters=rownames(rr))
```

```
barplot(rr$AUC[1:37], names=rr$parameters[1:37], col=colors_auc[1:37],ylim=c(0,1.2),
        horiz=F , las=2, ylab="Area under Density",cex.names=1.2,cex.axis=1.5)
barplot(rr$AUC[42:82], names=rr$parameters[42:82], col=colors_auc[42:82],ylim=c(0,1.2),
        horiz=F , las=2,ylab="Area under Density",cex.names=1.2,cex.axis=1.5)
##### to get the posterior summary of the marginal parameters #####
posterior.marginal<- matrix(rep(NA, length(marg.density)*3), ncol=3)
rownames(posterior.marginal)<- names(marg.density)
colnames(posterior.marginal)<- c("2.5%", "50%", "97.5%")
ignore.me<- union(grep("\\(Int", names(marg.density)), grep("prec",names(marg.density)))
comment<- rep("", length(marg.density))
for (i in 1:length(marg.density)) {tmp<- marg.density[[i]]; tmp2<- cumsum(tmp[,2])/sum(tmp[,2]
  posterior.marginal[i,]<-c(tmp[which(tmp2>0.025)[1]-1,1],tmp[which(tmp2>0.5)[1],1],
            tmp[which(tmp2>0.975)[1],1])
  vector<- posterior.marginal[i,]
  if( !(i%in%ignore.me) &&  (vector[1]<0 && vector[3]>0)){comment[i]<-"not sig. at 5%"}
  posterior.marginal[i,]<-as.numeric(formatC(posterior.marginal[i,],digits=4,format="f"))}
knitr::kable(cbind(posterior.marginal), row.names = TRUE, digits = 4, align = "rrrr", "html")
  kable_styling(bootstrap_options = "striped", full_width = FALSE, position = "left") %>%
  column_spec(3, bold = F)
#############################################################################################
## Parametric bootstrapping: for controlling overfitting (checking robustness of the DAG)
###                                  using JAGS                                        #####
#############################################################################################
m<- marg.new
env_aqhchoq.p<- cbind( m[["env_aqhchoq|(Intercept)"]], m[["env_aqhchoq|env_dst011"]],
                       m[["env_aqhchoq|demog_childst"]],m[["env_aqhchoq|demog_dpt1"]],
                       m[["env_aqhchoq|demog_dpt3"]],m[["env_aqhchoq|demog_literacy"]])
env_aqhchoq.prec.p<- cbind(m[["env_aqhchoq|precision"]])
env_dst011.p<- cbind( m[["env_dst011|(Intercept)"]], m[["env_dst011|env_elevation"]],
                      m[["env_dst011|infra_dmsp"]])
```

```
env_dst011.prec.p<- cbind(m[["env_dst011|precision"]])
env_dst190.p<- cbind(m[["env_dst190|(Intercept)"]], m[["env_dst190|env_dst011"]],
                     m[["env_dst190|env_elevation"]],m[["env_dst190|infra_water"]])
env_dst190.prec.p<- cbind(m[["env_dst190|precision"]])
env_elevation.p<- cbind(  m[["env_elevation|(Intercept)"]])
env_elevation.prec.p<- cbind(m[["env_elevation|precision"]])
infra_cluster2hcf.p<-cbind(m[["infra_cluster2hcf|(Intercept)"]], m[["infra_cluster2hcf|infra_d
            m[["infra_cluster2hcf|infra_water"]],m[["infra_cluster2hcf|demog_measles"]])
infra_cluster2hcf.prec.p<- cbind(m[["infra_cluster2hcf|precision"]])
infra_dmsp.p<- cbind( m[["infra_dmsp|(Intercept)"]],m[["infra_dmsp|demog_popdens"]],
                      m[["infra_dmsp|pop"]])
infra_dmsp.prec.p<- cbind(m[["infra_dmsp|precision"]])
demog_popdens.p<- cbind( m[["demog_popdens|(Intercept)"]])
demog_popdens.prec.p<- cbind(m[["demog_popdens|precision"]])
pop.p<- cbind( m[["pop|(Intercept)"]])
pop.prec.p<- cbind(m[["pop|precision"]])
infra_sanitation.p<-cbind(m[["infra_sanitation|(Intercept)"]],m[["infra_sanitation|env_dst011"
                    m[["infra_sanitation|infra_cluster2hcf"]])
infra_sanitation.prec.p<- cbind(m[["infra_sanitation|precision"]])
infra_water.p<- cbind( m[["infra_water|(Intercept)"]], m[["infra_water|env_elevation"]],
            m[["infra_water|demog_childst"]],m[["infra_water|demog_dpt1"]])
infra_water.prec.p<- cbind(m[["infra_water|precision"]])
demog_basicvaccs.p<- cbind( m[["demog_basicvaccs|(Intercept)"]],
    m[["demog_basicvaccs|env_elevation"]],m[["demog_basicvaccs|infra_water"]],
    m[["demog_basicvaccs|demog_dpt3"]], m[["demog_basicvaccs|demog_measles"]])
demog_basicvaccs.prec.p<- cbind(m[["demog_basicvaccs|precision"]])
demog_childst.p<-cbind(m[["demog_childst|(Intercept)"]],m[["demog_childst|env_dst011"]],
   m[["demog_childst|demog_popdens"]], m[["demog_childst|demog_dpt1"]])
demog_childst.prec.p<- cbind(m[["demog_childst|precision"]])
demog_dpt1.p<- cbind(m[["demog_dpt1|(Intercept)"]], m[["demog_dpt1|demog_measles"]])
```

```
demog_dpt1.prec.p<- cbind(m[["demog_dpt1|precision"]])
demog_dpt3.p<- cbind( m[["demog_dpt3|(Intercept)"]],m[["demog_dpt3|infra_cluster2hcf"]],
        m[["demog_dpt3|demog_dpt1"]])
demog_dpt3.prec.p<- cbind( m[["demog_dpt3|precision"]])
demog_literacy.p<-cbind(m[["demog_literacy|(Intercept)"]],
 m[["demog_literacy|demog_basicvaccs"]],m[["demog_literacy|demog_childst"]],
 m[["demog_literacy|demog_dpt3"]],m[["demog_literacy|demog_poverty"]])
demog_literacy.prec.p<- cbind(m[["demog_literacy|precision"]])
demog_literacy_inequality.p<-cbind( m[["demog_literacy_inequality|(Intercept)"]],
m[["demog_literacy_inequality|env_dst011"]], m[["demog_literacy_inequality|infra_water"]],
m[["demog_literacy_inequality|demog_literacy"]], m[["demog_literacy_inequality|demog_measles"]]
demog_literacy_inequality.prec.p<- cbind(m[["demog_literacy_inequality|precision"]])
demog_measles.p<- cbind( m[["demog_measles|(Intercept)"]])
demog_measles.prec.p<- cbind(m[["demog_measles|precision"]])
demog_poverty.p<- cbind( m[["demog_poverty|(Intercept)" ]],m[["demog_poverty|env_elevation"]],
            m[["demog_poverty|infra_dmsp"]],m[["demog_poverty|demog_measles"]])
demog_poverty.prec.p<- cbind(m[["demog_poverty|precision"]])
total_af.p<- cbind( m[[ "total_af|(Intercept)"]], m[[ "total_af|env_elevation"]],
 m[["total_af|demog_literacy"]], m[["total_af|demog_popdens"]],m[["total_af|total_scr"]])
total_scr.p<- cbind( m[["total_scr|(Intercept)"]])
dump(c("env_aqhchoq.p","env_aqhchoq.prec.p", "env_dst011.p","env_dst011.prec.p",
       "env_dst190.p","env_dst190.prec.p",  "env_elevation.p","env_elevation.prec.p",
       "infra_cluster2hcf.p","infra_cluster2hcf.prec.p","infra_dmsp.p","infra_dmsp.prec.p",
       "demog_popdens.p","demog_popdens.prec.p", "pop.p","pop.prec.p", "infra_sanitation.p",
       "infra_sanitation.prec.p","infra_water.p","infra_water.prec.p","demog_basicvaccs.p",
       "demog_basicvaccs.prec.p","demog_childst.p","demog_childst.prec.p","demog_dpt1.p",
       "demog_dpt1.prec.p","demog_dpt3.p","demog_dpt3.prec.p","demog_literacy.p",
       "demog_literacy.prec.p","demog_literacy_inequality.p",
       "demog_literacy_inequality.prec.p","demog_measles.p","demog_measles.prec.p",
       "demog_poverty.p","demog_poverty.prec.p","total_af.p","total_scr.p"),
```

```
     file=paste("DataForBoot_taf.R", sep=""))
## Data for JAGS
Jags.data<-list(
          'env_aqhchoq.p'=env_aqhchoq.p,'env_aqhchoq.prec.p'=env_aqhchoq.prec.p,
          'env_dst011.p'=env_dst011.p,'env_dst011.prec.p'=env_dst011.prec.p,
          'env_dst190.p'=env_dst190.p,'env_dst190.prec.p'=env_dst190.prec.p,
          'env_elevation.p'=env_elevation.p,'env_elevation.prec.p'=env_elevation.prec.p,
           'infra_cluster2hcf.p'=infra_cluster2hcf.p,
          'infra_cluster2hcf.prec.p'=infra_cluster2hcf.prec.p,
          'infra_dmsp.p'=infra_dmsp.p,'infra_dmsp.prec.p'=infra_dmsp.prec.p,
          'demog_popdens.p'=demog_popdens.p,'demog_popdens.prec.p'=demog_popdens.prec.p,
          'pop.p'=pop.p,'pop.prec.p'=pop.prec.p,'infra_sanitation.p'=infra_sanitation.p,
          'infra_sanitation.prec.p'=infra_sanitation.prec.p,'infra_water.p'=infra_water.p,
          'infra_water.prec.p'=infra_water.prec.p,'demog_basicvaccs.p'=demog_basicvaccs.p,
          'demog_basicvaccs.prec.p'=demog_basicvaccs.prec.p,
          'demog_childst.p'=demog_childst.p,
          'demog_childst.prec.p'=demog_childst.prec.p,'demog_dpt1.p'=demog_dpt1.p,
          'demog_dpt1.prec.p'=demog_dpt1.prec.p,
          'demog_dpt3.p'=demog_dpt3.p,'demog_dpt3.prec.p'=demog_dpt3.prec.p,
          'demog_literacy.p'=demog_literacy.p,'demog_literacy.prec.p'=demog_literacy.prec.p,
          'demog_literacy_inequality.p'=demog_literacy_inequality.p,
          'demog_literacy_inequality.prec.p'=demog_literacy_inequality.prec.p,
          'demog_measles.p'=demog_measles.p,'demog_measles.prec.p'=demog_measles.prec.p,
          'demog_poverty.p'=demog_poverty.p,'demog_poverty.prec.p'=demog_poverty.prec.p,
          'total_af.p'=total_af.p,'total_scr.p'=total_scr.p)
################################### Bootstrapping ############################
source("DataForBoot_taf.R")   # load data
JAGS_model_taf<-"model{
#- env_aqhchoq|env_dst011:demog_dpt1:demog_dpt3:demog_literacy
env_aqhchoq~dnorm(env_aqhchoq.mu,env_aqhchoq.prec);
```

```
env_aqhchoq.mu<- env_aqhchoq.c0 + env_aqhchoq.c1*demog_childst+env_aqhchoq.c2*demog_dpt1+
             env_aqhchoq.c3*demog_dpt3+env_aqhchoq.c4*demog_literacy;
 env_aqhchoq.M0~dcat(env_aqhchoq.p[ ,2]); env_aqhchoq.c0<-env_aqhchoq.p[env_aqhchoq.M0,1];
 env_aqhchoq.M1~dcat(env_aqhchoq.p[ ,4]);env_aqhchoq.c1<-env_aqhchoq.p[env_aqhchoq.M1,3];
 env_aqhchoq.M2~dcat(env_aqhchoq.p[ ,6]);env_aqhchoq.c2<-env_aqhchoq.p[env_aqhchoq.M2,5];
 env_aqhchoq.M3~dcat(env_aqhchoq.p[ ,8]);env_aqhchoq.c3<- env_aqhchoq.p[env_aqhchoq.M3,7];
 env_aqhchoq.M4~dcat(env_aqhchoq.p[ ,10]);env_aqhchoq.c4<- env_aqhchoq.p[env_aqhchoq.M4,9];
 env_aqhchoq.prec.M~dcat(env_aqhchoq.prec.p[ ,2]);
 env_aqhchoq.prec<-env_aqhchoq.prec.p[env_aqhchoq.prec.M,1];
 #- env_dst011|infra_dmsp:infra_sanitation
 env_dst011~dnorm(env_dst011.mu,env_dst011.prec);
 env_dst011.mu<- env_dst011.c0+env_dst011.c1*ienv_elevation+env_dst011.c2*infra_dsmp;
 env_dst011.M0~dcat(env_dst011.p[ ,2]);env_dst011.c0<- env_dst011.p[env_dst011.M0,1];
 env_dst011.M1~dcat(env_dst011.p[ ,4]);env_dst011.c1<- env_dst011.p[env_dst011.M1,3];
 env_dst011.M2~dcat(env_dst011.p[ ,6]);env_dst011.c2<- env_dst011.p[env_dst011.M2,5];
 env_dst011.prec.M~dcat(env_dst011.prec.p[ ,2]);
 env_dst011.prec<-env_dst011.prec.p[env_dst011.prec.M,1];
 #- env_dst190|env_dst011:env_elevation:infra_water
 env_dst190~dnorm(env_dst190.mu,env_dst190.prec);
 env_dst190.mu<- env_dst190.c0+env_dst190.c1*env_dst011+env_dst190.c2*env_elevation+
          env_dst190.c3*infra_water;
 env_dst190.M0~dcat(env_dst190.p[ ,2]); env_dst190.c0<- env_dst190.p[env_dst190.M0,1];
 env_dst190.M1~dcat(env_dst190.p[ ,4]);env_dst190.c1<- env_dst190.p[env_dst190.M1,3];
 env_dst190.M2~dcat(env_dst190.p[ ,6]);env_dst190.c2<- env_dst190.p[env_dst190.M2,5];
 env_dst190.M3~dcat(env_dst190.p[ ,8]);env_dst190.c3<-env_dst190.p[env_dst190.M3,7];
env_dst190.prec.M~dcat(env_dst190.prec.p[ ,2]);
env_dst190.prec<-env_dst190.prec.p[env_dst190.prec.M,1];
#- env_elevation|infra_dmsp:demog_basicvaccs:demog_poverty
env_elevation~dnorm(env_elevation.mu,env_elevation.prec);
env_elevation.mu<- env_elevation.c0;
```

```
env_elevation.M0~dcat(env_elevation.p[,2]);
env_elevation.c0<- env_elevation.p[env_elevation.M0,1];
env_elevation.prec.M~dcat(env_elevation.prec.p[ ,2]);
env_elevation.prec<-env_elevation.prec.p[env_elevation.prec.M,1];
 #- infra_cluster2hcf|infra_dmsp:demog_measles
  infra_cluster2hcf~dnorm(infra_cluster2hcf.mu,infra_cluster2hcf.prec);
 infra_cluster2hcf.mu<- infra_cluster2hcf.c0+infra_cluster2hcf.c1*infra_dmsp+
    infra_cluster2hcf.c2*infra_water+ infra_cluster2hcf.c3*demog_measles;
 infra_cluster2hcf.M0~dcat(infra_cluster2hcf.p[ ,2]);
 infra_cluster2hcf.c0<- infra_cluster2hcf.p[infra_cluster2hcf.M0,1];
 infra_cluster2hcf.M1~dcat(infra_cluster2hcf.p[ ,4]);
 infra_cluster2hcf.c1<- infra_cluster2hcf.p[infra_cluster2hcf.M1,3];
 infra_cluster2hcf.M2~dcat(infra_cluster2hcf.p[ ,6]);
 infra_cluster2hcf.c2<- infra_cluster2hcf.p[infra_cluster2hcf.M2,5];
 infra_cluster2hcf.M3~dcat(infra_cluster2hcf.p[ ,8]);
 infra_cluster2hcf.c3<- infra_cluster2hcf.p[infra_cluster2hcf.M3,7];
infra_cluster2hcf.prec.M~dcat(infra_cluster2hcf.prec.p[ ,2]);
 infra_cluster2hcf.prec<-infra_cluster2hcf.prec.p[infra_cluster2hcf.prec.M,1];
 #- infra_dmsp
 infra_dmsp~dnorm(infra_dmsp.mu,infra_dmsp.prec);
 infra_dmsp.mu<- infra_dmsp.c0+infra_dmsp.c1*demog_popdens+infra_dmsp.c2*pop;
 infra_dmsp.M0~dcat(infra_dmsp.p[ ,2]);infra_dmsp.c0<- infra_dmsp.p[infra_dmsp.M0,1];
 infra_dmsp.M1~dcat(infra_dmsp.p[ ,4]);infra_dmsp.c1<- infra_dmsp.p[infra_dmsp.M1,3];
 infra_dmsp.M2~dcat(infra_dmsp.p[ ,6]);infra_dmsp.c2<- infra_dmsp.p[infra_dmsp.M2,5];
 infra_dmsp.prec.M~dcat(infra_dmsp.prec.p[ ,2]);
 infra_dmsp.prec<-infra_dmsp.prec.p[infra_dmsp.prec.M,1];
 # demog_popdens
 demog_popdens~dnorm(demog_popdens.mu,demog_popdens.prec);
 demog_popdens.mu<- demog_popdens.c0; demog_popdens.M0~dcat(demog_popdens.p[ ,2]);
 demog_popdens.c0<- demog_popdens.p[demog_popdens.M0,1];
```

```
demog_popdens.prec.M~dcat(demog_popdens.prec.p[ ,2]);
 demog_popdens.prec<-demog_popdens.prec.p[demog_popdens.prec.M,1];
 ############# pop|
 pop~dnorm(pop.mu,pop.prec); pop.mu<- pop.c0;  pop.M0~dcat(pop.p[ ,2]);
pop.c0<-pop.p[pop.M0,1];pop.prec.M~dcat(pop.prec.p[ ,2]);pop.prec<-pop.prec.p[pop.prec.M,1];
 # infra_sanitation|infra_cluster2hcf
  infra_sanitation~dnorm(infra_sanitation.mu,infra_sanitation.prec);
infra_sanitation.mu<- infra_sanitation.c0+infra_sanitation.c1*env_dst011+
     infra_sanitation.c2*infra_cluster2hcf;
 infra_sanitation.M0~dcat(infra_sanitation.p[ ,2]);
 infra_sanitation.c0<- infra_sanitation.p[infra_sanitation.M0,1];
 infra_sanitation.M1~dcat(infra_sanitation.p[ ,4]);
 infra_sanitation.c1<- infra_sanitation.p[infra_sanitation.M1,3];
  infra_sanitation.M2~dcat(infra_sanitation.p[ ,6]);
 infra_sanitation.c2<- infra_sanitation.p[infra_sanitation.M2,5];
infra_sanitation.prec.M~dcat(infra_sanitation.prec.p[ ,2]);
 infra_sanitation.prec<-infra_sanitation.prec.p[infra_sanitation.prec.M,1];
 # infra_water
 infra_water~dnorm(infra_water.mu,infra_water.prec);
 infra_water.mu<- infra_water.c0+infra_water.c1*env_elevation+
 infra_water.c2*demog_childst+infra_water.c3*demog_dpt1;
 infra_water.M0~dcat(infra_water.p[ ,2]);infra_water.c0<- infra_water.p[infra_water.M0,1];
 infra_water.M1~dcat(infra_water.p[ ,4]); infra_water.c1<- infra_water.p[infra_water.M1,3];
 infra_water.M2~dcat(infra_water.p[ ,6]);infra_water.c2<- infra_water.p[infra_water.M2,5];
 infra_water.M3~dcat(infra_water.p[ ,8]);infra_water.c3<- infra_water.p[infra_water.M3,7];
infra_water.prec.M~dcat(infra_water.prec.p[ ,2]);
 infra_water.prec<-infra_water.prec.p[infra_water.prec.M,1];
 # demog_basicvaccs
 demog_basicvaccs~dnorm(demog_basicvaccs.mu,demog_basicvaccs.prec);
 demog_basicvaccs.mu<- demog_basicvaccs.c0+demog_basicvaccs.c1*env_elevation+
```

```
  demog_basicvaccs.c2*infra_water+demog_basicvaccs.c3*demog_dpt3+demog_basicvaccs.c4*demog_mea
 demog_basicvaccs.M0~dcat(demog_basicvaccs.p[ ,2]);
 demog_basicvaccs.c0<- demog_basicvaccs.p[demog_basicvaccs.M0,1];
 demog_basicvaccs.M1~dcat(demog_basicvaccs.p[ ,4]);
 demog_basicvaccs.c1<- demog_basicvaccs.p[demog_basicvaccs.M1,3];
 demog_basicvaccs.M2~dcat(demog_basicvaccs.p[ ,6]);
 demog_basicvaccs.c2<- demog_basicvaccs.p[demog_basicvaccs.M2,5];
 demog_basicvaccs.M3~dcat(demog_basicvaccs.p[ ,8]);
 demog_basicvaccs.c3<- demog_basicvaccs.p[demog_basicvaccs.M3,7];
 demog_basicvaccs.M4~dcat(demog_basicvaccs.p[ ,10]);
 demog_basicvaccs.c4<- demog_basicvaccs.p[demog_basicvaccs.M4,9];
demog_basicvaccs.prec.M~dcat(demog_basicvaccs.prec.p[ ,2]);
  demog_basicvaccs.prec<-demog_basicvaccs.prec.p[demog_basicvaccs.prec.M,1];
 # demog_childst
demog_childst~dnorm(demog_childst.mu,demog_childst.prec);
demog_childst.mu<-demog_childst.c0+demog_childst.c1*env_dst011+
demog_childst.c2*demog_popdens+demog_childst.c1*demog_dpt1;
demog_childst.M0~dcat(demog_childst.p[,2]);
demog_childst.c0<-demog_childst.p[demog_childst.M0,1];
demog_childst.M1~dcat(demog_childst.p[,4]);
demog_childst.c1<-demog_childst.p[demog_childst.M1,3];
demog_childst.M2~dcat(demog_childst.p[,6]);
demog_childst.c2<-demog_childst.p[demog_childst.M2,5];
demog_childst.M3~dcat(demog_childst.p[,6]);
demog_childst.c3<-demog_childst.p[demog_childst.M3,7];
demog_childst.prec.M~dcat(demog_childst.prec.p[ ,2]);
demog_childst.prec<-demog_childst.prec.p[demog_childst.prec.M,1];
  # demog_dpt1
demog_dpt1~dnorm(demog_dpt1.mu,demog_dpt1.prec);
demog_dpt1.mu<- demog_dpt1.c0+demog_dpt1.c1*demog_measles;
```

```
demog_dpt1.M0~dcat(demog_dpt1.p[ ,2]); demog_dpt1.c0<- demog_dpt1.p[demog_dpt1.M0,1];
demog_dpt1.M1~dcat(demog_dpt1.p[ ,4]); demog_dpt1.c1<- demog_dpt1.p[demog_dpt1.M1,3];
demog_dpt1.prec.M~dcat(demog_dpt1.prec.p[,2]);
demog_dpt1.prec<-demog_dpt1.prec.p[demog_dpt1.prec.M,1];
  # demog_dpt3
demog_dpt3~dnorm(demog_dpt3.mu,demog_dpt3.prec);
demog_dpt3.mu<- demog_dpt3.c0+demog_dpt3.c1*infra_cluster2hcf+demog_dpt3.c2*demog_dpt1;
demog_dpt3.M0~dcat(demog_dpt3.p[ ,2]); demog_dpt3.c0<- demog_dpt3.p[demog_dpt3.M0,1];
demog_dpt3.M1~dcat(demog_dpt3.p[ ,4]); demog_dpt3.c1<- demog_dpt3.p[demog_dpt3.M1,3];
demog_dpt3.M2~dcat(demog_dpt3.p[ ,6]);   demog_dpt3.c2<- demog_dpt3.p[demog_dpt3.M2,5];
demog_dpt3.prec.M~dcat(demog_dpt3.prec.p[,2]);
demog_dpt3.prec<-demog_dpt3.prec.p[demog_dpt3.prec.M,1];
       # demog_literacy
demog_literacy~dnorm(demog_literacy.mu,demog_literacy.prec);
demog_literacy.mu<- demog_literacy.c0+demog_literacy.c1*demog_basicvaccs+
demog_literacy.c2*demog_childst+demog_literacy.c3*demog_dpt3+demog_literacy.c4*demog_poverty;
demog_literacy.M0~dcat(demog_literacy.p[,2]);
demog_literacy.c0<-demog_literacy.p[demog_literacy.M0,1];
demog_literacy.M1~dcat(demog_literacy.p[,2]);
demog_literacy.c1<-demog_literacy.p[demog_literacy.M1,3];
demog_literacy.M2~dcat(demog_literacy.p[,4]);
demog_literacy.c2<-demog_literacy.p[demog_literacy.M2,5];
demog_literacy.M3~dcat(demog_literacy.p[,6]);
demog_literacy.c3<-demog_literacy.p[demog_literacy.M3,7];
demog_literacy.M4~dcat(demog_literacy.p[,8]);
demog_literacy.c4<-demog_literacy.p[demog_literacy.M4,9];
demog_literacy.prec.M~dcat(demog_literacy.prec.p[,2]);
demog_literacy.prec<-demog_literacy.prec.p[demog_literacy.prec.M,1];
# demog_literacy_inequality
demog_literacy_inequality~dnorm(demog_literacy_inequality.mu,demog_literacy_inequality.prec);
```

```
demog_literacy_inequality.mu<- demog_literacy_inequality.c0+
   demog_literacy_inequality.c1*env_dst011+demog_literacy_inequality.c2*infra_water+
   demog_literacy_inequality.c3*demog_literacy+demog_literacy_inequality.c3*demog_measles ;
demog_literacy_inequality.M0~dcat(demog_literacy_inequality.p[ ,2]);
demog_literacy_inequality.c0<- demog_literacy_inequality.p[demog_literacy_inequality.M0,1];
demog_literacy_inequality.M1~dcat(demog_literacy_inequality.p[ ,4]);
demog_literacy_inequality.c1<-demog_literacy_inequality.p[demog_literacy_inequality.M1,3];
demog_literacy_inequality.M2~dcat(demog_literacy_inequality.p[ ,6]);
demog_literacy_inequality.c2<- demog_literacy_inequality.p[demog_literacy_inequality.M2,5];
demog_literacy_inequality.M3~dcat(demog_literacy_inequality.p[ ,8]);
demog_literacy_inequality.c3<- demog_literacy_inequality.p[demog_literacy_inequality.M3,7];
demog_literacy_inequality.M4~dcat(demog_literacy_inequality.p[ ,10]);
demog_literacy_inequality.c4<- demog_literacy_inequality.p[demog_literacy_inequality.M4,9];
demog_literacy_inequality.prec.M~dcat(demog_literacy_inequality.prec.p[ ,2]);
demog_literacy_inequality.prec<-
     demog_literacy_inequality.prec.p[demog_literacy_inequality.prec.M,1];
demog_measles~dnorm(demog_measles.mu,demog_measles.prec);
demog_measles.mu<- demog_measles.c0
demog_measles.M0~dcat(demog_measles.p[,2]);demog_measles.c0<-demog_measles.p[demog_measles.M0,
demog_measles.prec.M~dcat(demog_measles.prec.p[ ,2]);
demog_measles.prec<-demog_measles.prec.p[demog_measles.prec.M,1];
demog_poverty~dnorm(demog_poverty.mu,demog_poverty.prec);
demog_poverty.mu<- demog_poverty.c0+demog_poverty.c1*env_elevation+
                          demog_poverty.c2*infra_dmsp+demog_poverty.c3*demog_measles;
demog_poverty.M0~dcat(demog_poverty.p[,2]);
demog_poverty.c0<-demog_poverty.p[demog_poverty.M0,1];
demog_poverty.M1~dcat(demog_poverty.p[,4]);
demog_poverty.c1<-demog_poverty.p[demog_poverty.M1,3];
demog_poverty.M2~dcat(demog_poverty.p[,6]);
demog_poverty.c2<-demog_poverty.p[demog_poverty.M2,5];
```

```
demog_poverty.M3~dcat(demog_poverty.p[,8]);
demog_poverty.c3<-demog_poverty.p[demog_poverty.M3,7];
demog_poverty.prec.M~dcat(demog_poverty.prec.p[,2]);
demog_poverty.prec<-demog_poverty.prec.p[demog_poverty.prec.M,1];
#- total_af|env_elevation:demog_popdens:demog_literacy:total_scr
total_af~dpois(ptotal_af);
log(ptotal_af)<- total_af.c0 + total_af.c1*env_elevation+total_af.c2*demog_popdens+
                      total_af.c3*demog_literacy+total_af.c4*total_scr;
total_af.M0~dcat(total_af.p[ ,2]);total_af.c0<- total_af.p[total_af.M0,1];
total_af.M1~dcat(total_af.p[ ,4]);total_af.c1<- total_af.p[total_af.M1,3];
total_af.M2~dcat(total_af.p[ ,6]);total_af.c2<- total_af.p[total_af.M2,5];
total_af.M3~dcat(total_af.p[ ,8]); total_af.c3<- total_af.p[total_af.M3,7];
total_af.M4~dcat(total_af.p[ ,10]);total_af.c4<-total_af.p[total_af.M4,9];
total_scr~dpois(ptotal_scr);log(ptotal_scr)<- total_scr.c0 ;
total_scr.M0~dcat(total_scr.p[ ,2]);total_scr.c0<- total_scr.p[total_scr.M0,1];
}"
variable<- colnames(data);n<- sample(1:100000, 25000);start.time<-Sys.time()
for (i in 1:length(n)) {print(paste("Running simulation", i))
  # random initial values
  init<- list(".RNG.name"="base::Mersenne-Twister", ".RNG.seed"=n[i])
JAGS_total_af<- jags.model(file="JAGS_Model_taf",data =Jags.data,inits=init,n.chains=1,
 n.adapt=500)
  # to run more iterations
  update(JAGS_total_af, 100000)
  # sample data (same size as original: 300)
  boot.sample<- coda.samples(JAGS_total_af,
        c("env_aqhchoq","env_aqhchoq.prec", "env_dst011","env_dst011.prec",
        "env_dst190","env_dst190.prec","env_elevation","env_elevation.prec",
        "infra_cluster2hcf","infra_cluster2hcf.prec","infra_dmsp","infra_dmsp.prec",
        "demog_popdens","demog_popdens.prec", "pop","pop.prec",
```

```
        "infra_sanitation","infra_sanitation.prec","infra_water","infra_water.prec",
        "demog_basicvaccs","demog_basicvaccs.prec","demog_childst",
        "demog_childst.prec","demog_dpt1","demog_dpt1.prec","demog_dpt3","demog_dpt3.prec",
        "demog_literacy","demog_literacy.prec","demog_literacy_inequality",
        "demog_literacy_inequality.prec","demog_measles","demog_measles.prec",
        "demog_poverty","demog_poverty.prec","total_af","total_scr"),n.iter= 3000 , thin =10)
  # to create a data frame in the same shape as the original dataset
  boot.data<- as.data.frame(as.matrix(boot.sample))
  boot.data<- boot.data[, variable]
  # Build a cache of all local computations
  mycache.boot<- buildScoreCache(data.df = boot.data,data.dists = dists,max.parents=max.par,
              dag.banned=~total_scr|., method = "bayes")
  # To construct a dag based on the bootstrap sample using EXACT SEARCH algorithm
  dag.boot<- mostProbable(score.cache = mycache.boot,)
  fabn.boot<- fitAbn(object = dag.boot,data.df=boot.data,data.dists=dists,score = "mlik")
  boot.dags[[i]]<- dag.boot$dag};end.time<-Sys.time(); end.time-start.time
#To store the bootstrap sample
save(boot.dags, file = sprintf('boot.dags.RData'))
## count total number of arcs in each dag
arc.freq<-sapply(dags, sum);barplot(table(arcs.freq))
# Count how many times each arc appear in the bootstrap data
total.dag<-matrix(rep(0,dim(bestdag)[2]^2),ncol=dim(bestdag)[2]);
colnames(total.dag)<-rownames(total.dag)<-colnames(bestdag);
for(i in 1:length(boot.dags)){
  if(sum(boot.dags[[i]])>0){total.dag<-total.dag+boot.dags[[i]];}}
total.dag<-total.dag*bestdag;total.dag;
f<-function(val,limit){if(val<limit){return(0);}  else {return(1);}}
bestdag.trim<-apply(total.dag,c(1,2),FUN=f,limit=N/2);bestdag.trim.nodir<-bestdag;
bestdag.trim.nodir[,]<-0;child<-NULL;parent<-NULL;
for(i in 1:dim(total.dag)[1]){
```

```
for(j in 1:dim(total.dag)[2]){ if(i>j){
 if(total.dag[i,j]>total.dag[j,i]){m.i<-i;m.j<-j;} else {m.i<-j;m.j<-i;}
  if(total.dag[i,j]+total.dag[j,i]>N/2){bestdag.trim.nodir[m.i,m.j]<-1;}}}}
all.equal(bestdag.trim, bestdag.trim.nodir)
## number of arcs in Pruned (trimmed) DAGS
sum(bestdag.trim$dag) #  arcs, as in original model
####Best dag: bestdag.trim
#########################################################
##### The same procedure can be apply for bacteriological diagnosed TB cases
#########################################################
```