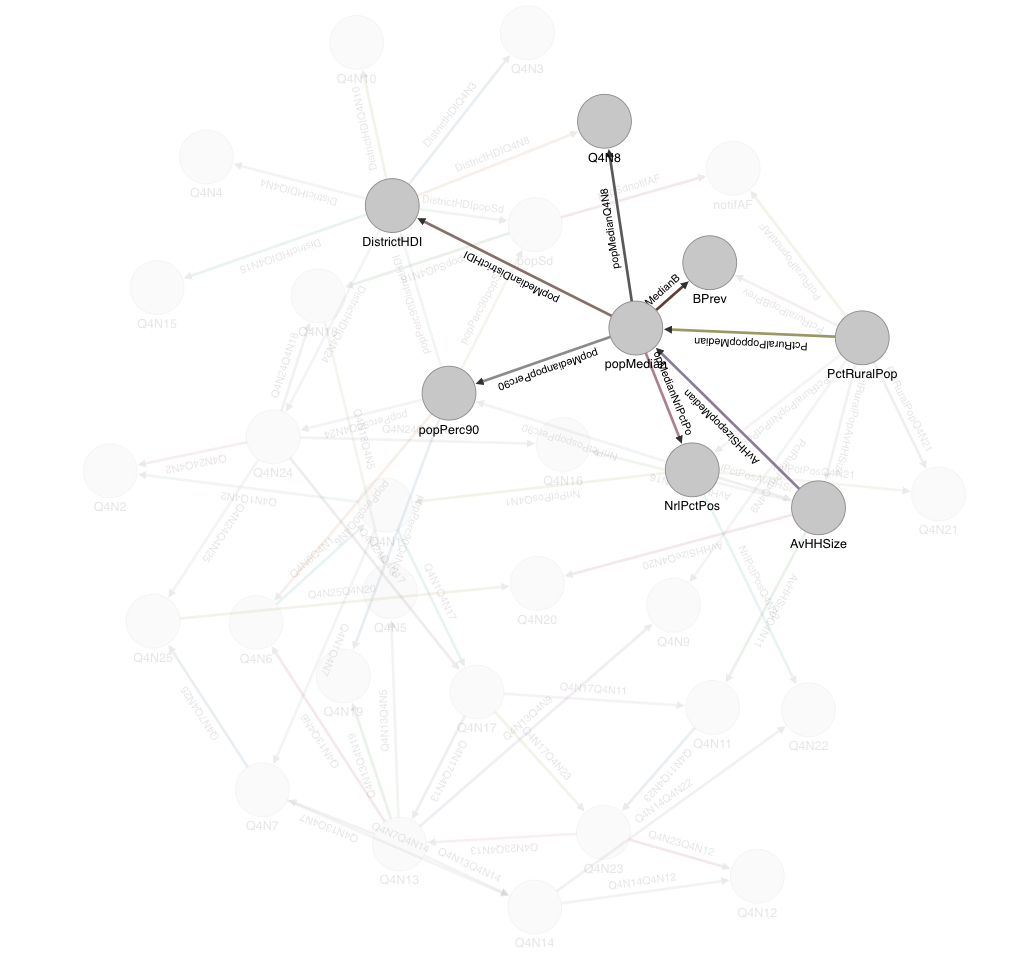
**Pakistan TB Hackathon**

**Submission**

Team: United4TB

Team members: Zhi Zhen Qin, Farihah Malik, Robert Williamson, Yumna Moosa, Matthys Potgieter, Anet Potgieter, Vincent Meurrens



1. **Background**

National population-based prevalence surveys provide a direct measurement of the absolute burden of disease and have been the gold standard in estimating the prevalence and trend for diseases such as tuberculosis (TB)1,2. However, a prevalence survey requires enormous technical and financial resources and it is not feasible to conduct such surveys frequently. Estimating the burden of disease in the absence of prevalence survey data is challenging, especially in low resource, high burden countries that do not have robust high-quality routine surveillance systems in place. We set out to develop a predictive model for sub-national TB burden in Pakistan. This paper describes the methods we used and results for the district-wise bacteriologically positive TB prevalence among adults (>=15 years) in Pakistan for 2018.

1. **Methods**
   1. **Justification for using Bayesian Network Modelling**

Bayesian networks provide a powerful Artificial Intelligence (AI) technology for probabilistic reasoning and statistical inference that can be used to reason with uncertainty in complex environments – for example in this project where the underlying data (surveillance, socio-economic and spatial-environmental datasets) with multiple variables that have many hidden relationships, having large amounts of natural variation, measurement errors, or missing values3-6. We used a machine learning (ML) engine to perform Bayesian learning (section 2.2) and Bayesian reasoning (section 2.3) that allows inference and predictive what-if queries on newly observed variables based on prior learning. A Bayesian network contains a set of predictor variables (represented as nodes), regardless of previously known associations with an outcome variable3-6. The links between the nodes represent informational or causal dependencies among the variables. The dependencies are given in terms of conditional probabilities of states that a node can have given the values of the parent nodes3-6. Each probability reflects a degree of belief. Degree of belief encodes causal dependencies. Degree of belief in any cause of a given effect is increased when the effect is observed, but then decreases when some other cause is found to be responsible for the observed effect. Causal dependencies can be derived from the knowledge of domain experts, or by mining the structure of the model from data by using unsupervised learning3-6.

In our model, we included a diverse range of data sources that are known to be associated with TB prevalence7-13. The data are grouped into the following four groups (details described in Table 1):

* + 1. Routine disease surveillance data
    2. Prevalence survey data
    3. Socio-economic data
    4. Spatial-Temporal Environmental Data

*Table 1 Module Input*

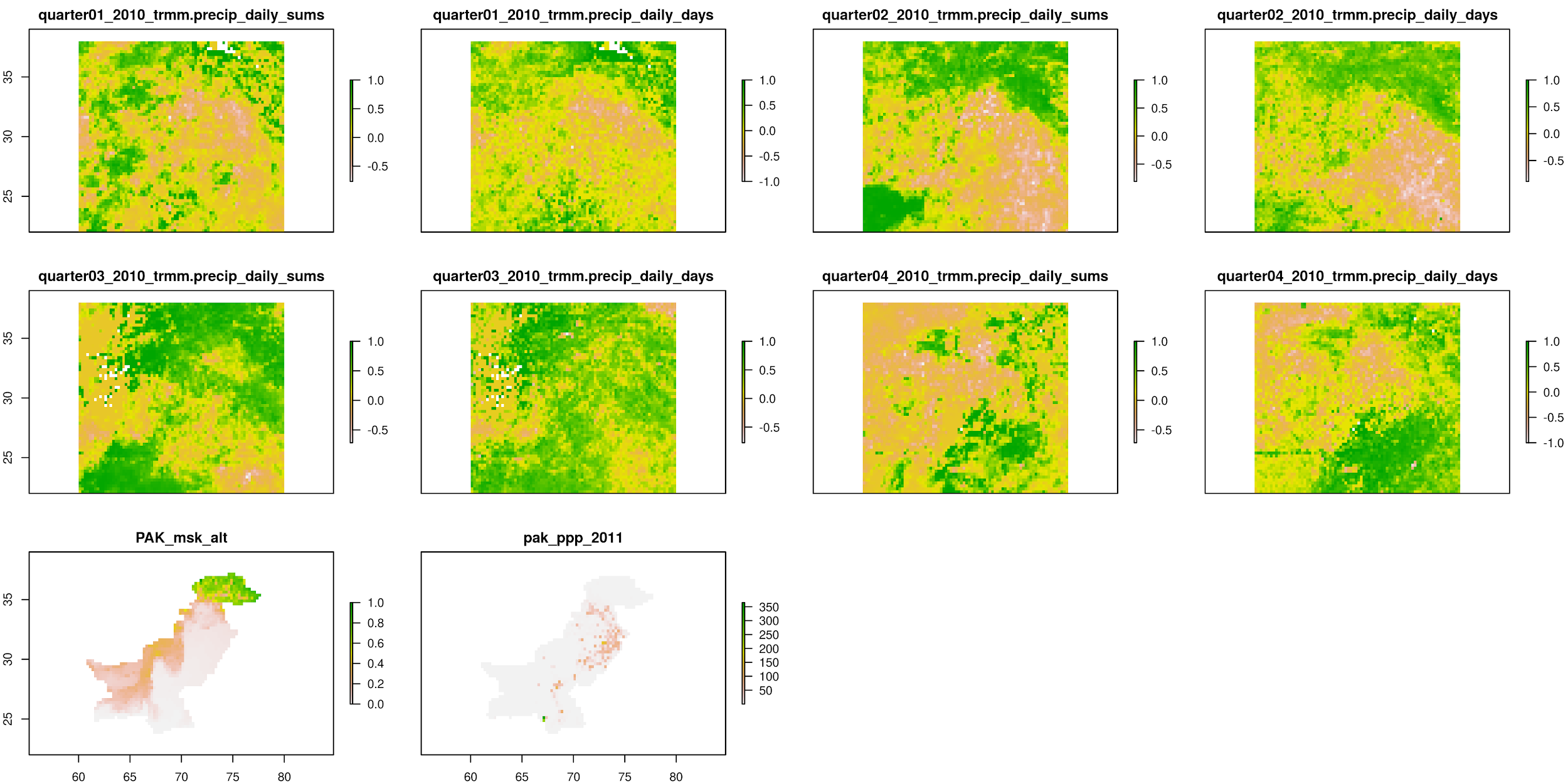
|  |  |  |
| --- | --- | --- |
| **Variable name** | **Definition** | |
| **2010-11 Pakistan TB Prevalence Survey Variables** | | |
| *BPrev* | Measured district-wise Bac+ prevalence from 2010/2011 prevalence survey, measured in number of cases per 100 000 population | |
| **ROUTINE DISEASE SURVEILLANCE VARIABLES** | | |
| *notifAF, notifAFQ1 - Q4* | District-wise TB notification (all forms) from notification (TB-07 form) | |
| *notifBposQ1 - Q4* | District-wise TB notification (Bac+) from notification (TB-07 form) | |
| *NrlPctPos* | Yearly district-wise smear positive rate among people tested for smear according to NRL data | |
| **SOCIO-ECONOMIC VARIABLES** | | |
| *popMedian* | Median population density in a district, obtained from ~100 m resolution WorldPop estimates 14 | |
| *popPerc90* | 90% percentile population density in a district - WorldPop14 | |
| *popSd* | Standard deviation of population density in a district - WorldPop14 | |
| *AvHHSize* | District-wise average household size - calculated from 2017 census results | |
| *PctRuralPop* | Percentage of population is rural (Census) | |
| *ruralGrowthRate* | District-wise rural growth rate obtained from the 2017 census results | |
| *ruralSexRatio2017* | District-wise sex ratio in rural areas obtained from 2017 census results | |
| *totalGrowthRate* | District-wise total growth rate obtained from 2017 census results | |
| *totalSexRatio2017* | District-wise total sex ratio in rural and urban areas in 2017 | |
| *urbanGrowthRate* | District-wise urban growth rate in 2017 (Census) | |
| *urbanSexRatio2017* | District-wise sex ratio in urban areas in 2017 (Census) | |
| *LogGrossNationalIncomePerCapita* | [[1]](#footnote-1)Yearly province-level log of gross national income per capita - obtained from Global Data Lab (GDL)15 | |
| *LifeExpectancy* | Yearly province-level life expectancy, GDL15 | |
| *ExpectedYearsSchooling* | Yearly province-level expected years of schooling, GDL15 | |
| *MeanYearsSchooling* | Yearly province-level mean years of schooling, GDL15 | |
| *SubNationalHdi* | Yearly province-wise human development index (HDI), GDL15 | |
| *DistrictHDI* | [[2]](#footnote-2)District-wise HDI, calculated by UNDP16 | |
| **SPATIAL-TEMPORAL ENVIRONMENTAL FEATURES** | | |
| *Q4N1-Q4N25* | | Quarterly consecutive precipitation (proxy for consecutive indoor days), other precipitation metrics, topographic height and population density recorded at each 25 km pixel covering Pakistan11. (Details described in Section 2.1.1) |

* + 1. **Spatial-Temporal Environmental Features**

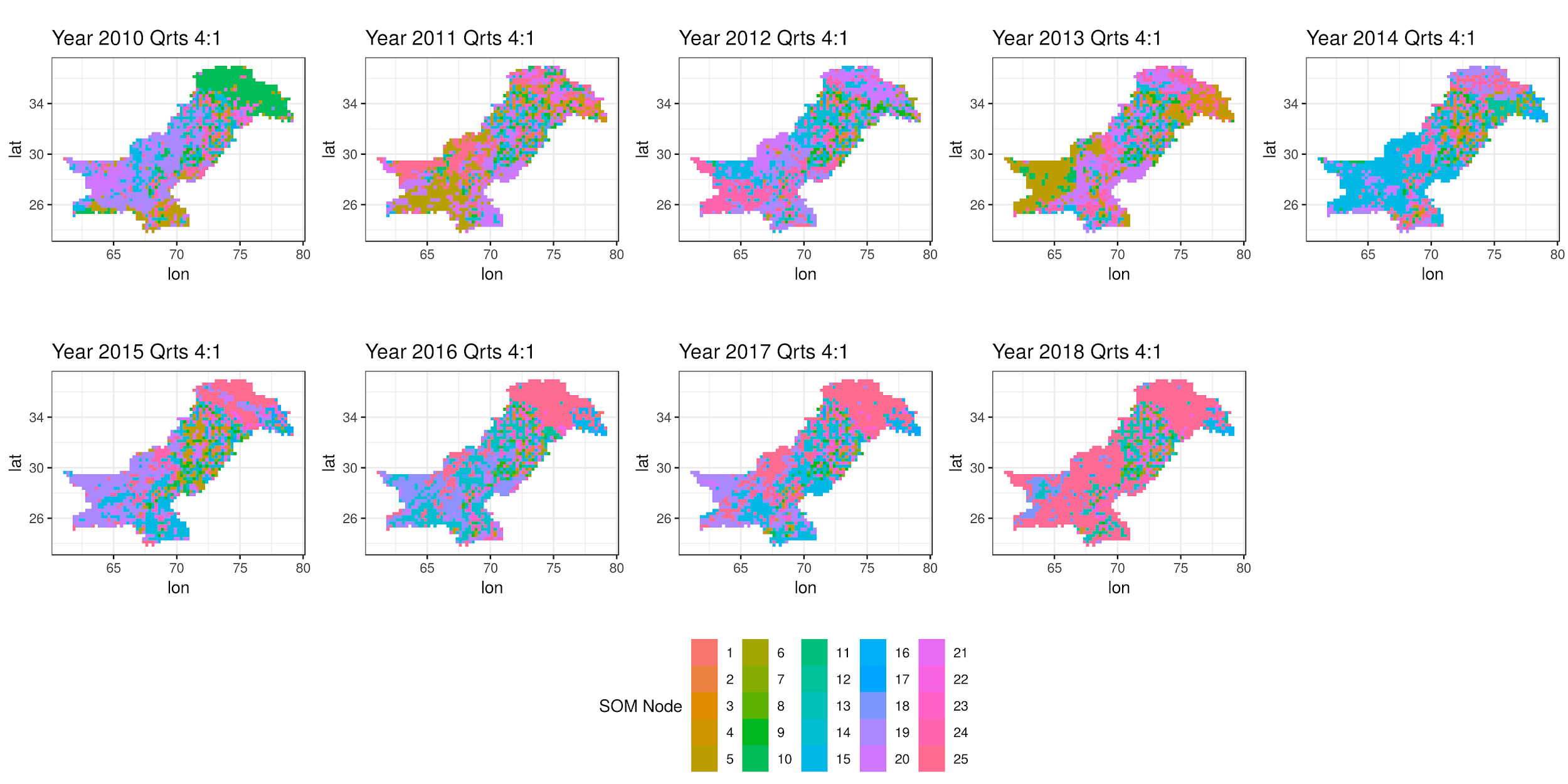
TB transmission often occurs through person-to-person within a household or small community because prolonged duration of contact is typically required for infection to occur17, as well as through reactivation of latent infection in a group of people with shared risk factors8,17,18. Spatial-environmental analysis has been promoted for targeted TB control and intensified use of existing TB control tools12,19. We used self-organizing maps (SOM) to extract patterns or features from multiple sources of spatial-environmental data that have been shown to influence incidence of tuberculosis. We included quarterly precipitation metrics for four consecutive quarters, topographic height and population density recorded at each 25 km pixel covering Pakistan7,11. The precipitation metrics are comprised of total quarterly rainfall and mean number of consecutive rainfall days, where the latter is used as a proxy for consecutive indoor days respectively. The indoor days suggests the time people spend in proximity, which is associated with TB transmission17. In addition to precipitation, topographic height and population density were included to differentiate where regional scale rainfall might differ in its influence on tuberculosis on a smaller scale20. Thus, a region of consistent precipitation metrics may be mapped to different SOM features or nodes according to height above sea-level, population density or both. Additional meteorological variables that similarly have been shown to have a significant impact on incidence of tuberculosis, such as temperature, relative humidity, atmospheric pressure and wind9 can be included in future SOMs.

SOMs use unsupervised simple competitive learning based on Euclidean distance between nodes and the input vector to incrementally adjust the weights of the nearest matching node and its neighbourhood. In this way, nodes in close proximity represent frequent similar yet distinctive patterns while allowing for unusual patterns to be accounted for.21,22. We created a SOM with a user-determined number of nodes (n=25) for each of the four annual quarters (training data are from years 2010 to 2017) so that for a given quarter each pixel vector maps to the nearest matching node in the corresponding SOM and is encoded by the node identifier. Consecutive quarters were used to capture seasonal patterns in rainfall as evidence suggests tuberculosis notification too can be seasonal10 and so might be linked to meteorological seasons. We decided on 25 nodes for each quarterly SOM thus reducing the multi-dimensional input space to 25 possible feature vectors. A winning node for each pixel for a given quarter thus encapsulates the four consecutive quarterly precipitation variability, topographic height and population density and the occurrence of that node across Pakistan identifies areas with similar characteristics.

*Figure 1 An example of the precipitation metrics, topographic height and population density for the four quarters of 2010 used to train the SOM for quarter 4. Each SOM input vector from each pixel has a dimension of 10*



*Figure 2 The SOM output mapping for quarter 4 of each year. Each SOM node is a feature that includes information on the current and previous three quarters precipitation and also topographic height and population density*



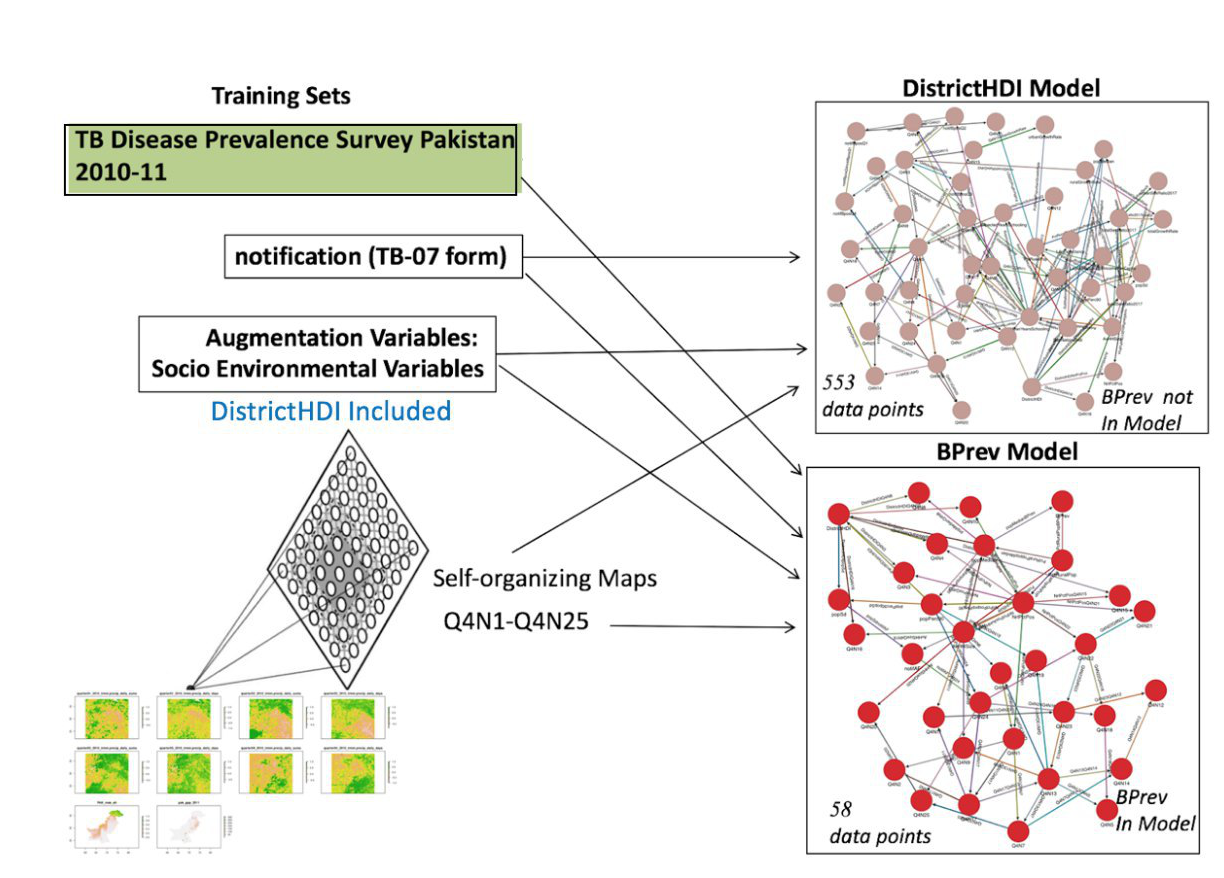
* 1. **Bayesian Learning**

Bayesian learning can be viewed as finding the local maxima on the likelihood surface defined by the Bayesian network variables. Assume that the network must be trained from D, a set of data cases D1, ..., Dm generated independently from some underlying distribution. In each data case, values are given for some subset of the variables; this subset may differ from case to case – in other words, the data can be incomplete. During Bayesian Learning, the parameters ω of the conditional probability matrices that best model the data are calculated. Our ML engine uses unsupervised Bayesian Learning to mine hidden relationships from the training sets using a *Hybrid Genetic Algorithm* (HGA)23,24.

Using unsupervised Bayesian learning we mined two models, namely:

* DistrictHDI model - used 553 data points in training set (trained using known district HDI for the year 2011-2015 and then to predict district-wise HDI for 2016, 2017 and 2018)
* BPrev model - used 58 data points in training set (trained using known 2011 district-wise Bac+ prevalence to predict 2018 district-wise Bac+ prevalence)

*Figure 3 The Bayesian Learning pipeline.*



*Figure 4 The direct relationships in the Bayesian Networks above.*

|  |  |
| --- | --- |
| DistrictHDI direct dependencies in DistrictHDI Bayesian Network mined from 553 data points | BPrev, DistrictHDI direct dependencies in BPrev Bayesian Network mined from limited dataset (58 points) |

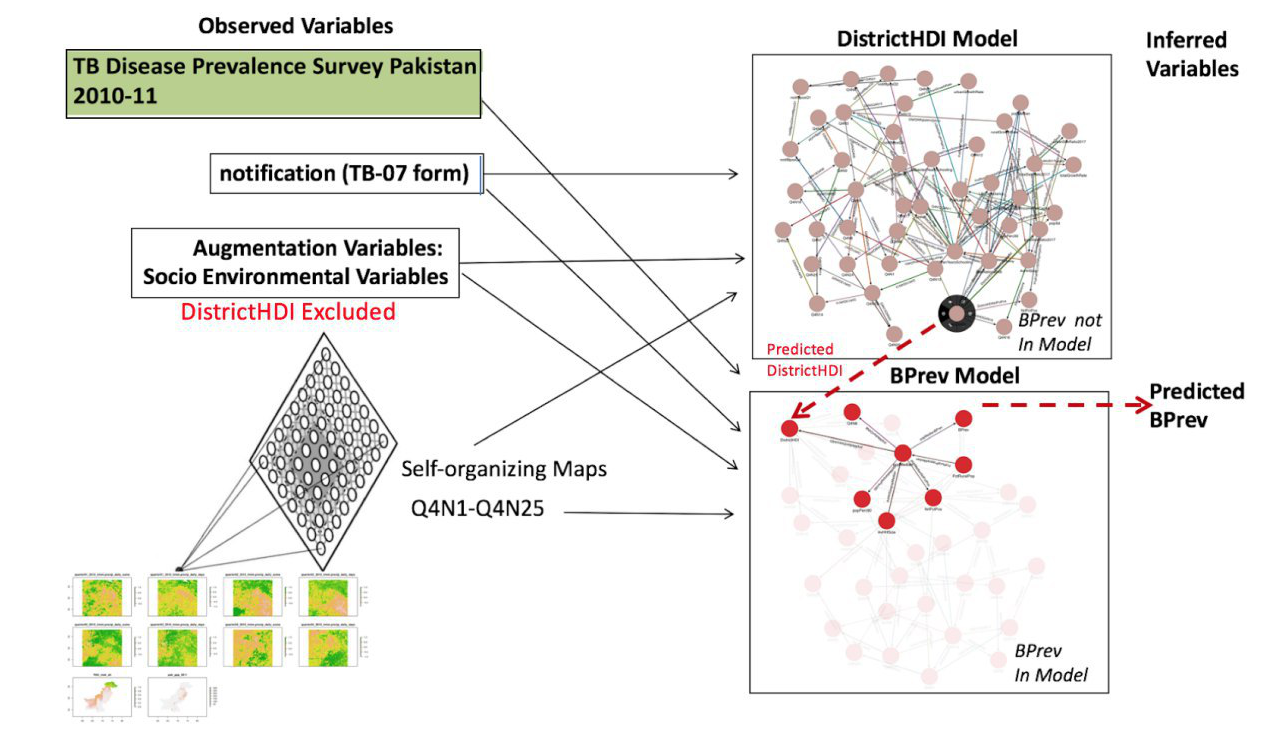
* 1. **Bayesian Reasoning**

Bayesian Belief Propagation is the process of finding the most probable explanation (MPE) in the presence of evidence. Following network structure generation, Bayesian inference can be performed to predict unknown variables in the network based on the states of observed nodes using Bayesian reasoning techniques. An algorithm, patented in South Africa and published by the World Intellectual Property Foundation (WIPO), is used to resolve queries in cyclical networks.25

Our Bayesian engine used the DistrictHDI model with augmented variables to predict DistrictHDI, which was used as input to the BPrev model mined from the model trained with the limited BPrev training set.

The Bayesian Reasoning pipeline is illustrated in Figure 4. The observed variables for the DistrictHDI model excludes the DistrictHDI variable. The first model predicts the DistrictHDI, which is then added to the observed variables that is presented to the BPrev model. The BPrev model then uses the observed variables, together with the predicted DistrictHDI, to infer BPrev.

*Figure 4 Bayesian reasoning pipeline*



* 1. **Credible Intervals**

Model output is obtained as inter-percentile ranges with an associated probability for each range - summing to 1 over all ranges. The probability density distribution is estimated by calculating the height of the probability distribution within each range. By discretizing the probability values over all ranges, we estimate the position of the 2.5th and 97.5th the percentiles to obtain the 95% **Credible Interval** for model predictions. Results for the optimum value for **BPrev** are in the **BPrev.Mid**, with the 95% **Credible Interval** specified by **BPrev.Low** (2.5%) and **BPrev.High** (97.5%) percentiles. **DistrictHDI** results are reported similarly.

* 1. **Training and Validation**

All source code is provided in the “united4TB.zip” file. Due to the inclusion of provided TB notification and other data under non-disclosure agreement (NDA), the GIT repository has not been made public. The open source pipeline that was developed for this project is replicable. It integrates the collectors consuming to the Pakistan training sets with the Bayesian engine. It uses Apache NiFi (<https://nifi.apache.org/>) - an open source software for automating and managing the flow of data between collectors and the Bayesian engine. Apache Nifi is a powerful and reliable system to process and distribute data. It provides a web-based User Interface for creating, monitoring, and controlling data flows. The Bayesian Models are visualised in OrientDB (<https://orientdb.com/>) - an open source database platform that can integrate into different languages such as java, python, R and so forth.

The following scripts were written to train and validate the model (these are placed in “united4tb/08 R script” folder):

1. **create\_tables.R -** To upload the combined data set to the Bayesian engine - a combined table for all districts, and a table for each district to be predicted from the prevalence survey, with the data for that district left out.
2. **xvalidation.R -** To predict district level prevalence using leave-one-out cross-validation, and querying the model with all available data for each district (*predictions/xvalidation.csv*). This is the final cross-validation output as part of the submission criteria.
3. **xvalidation\_som.R -** To predict district level prevalence using leave-one-out cross-validation, and querying the model with only SOMS + notifAF for each district (*predictions/xvalidation\_soms.csv*).
4. **xvalidation\_notifAF.R -** To predict district level prevalence using leave-one-out cross validation and querying the model with only notifAF for each district (*predictions/xvalidation\_notifAF.csv*).
5. **benchmark.R**  - generates **Root Mean Squared Error** and **Mean Squared Error** for the output of each cross-validation run, so that their performance can be compared.
6. **TBBurden.R** - generates sub-national Bac+ TB burden estimates at district level using the model trained on all districts, using all available data - including SOMS, TB notifications, population density and social indicators. (*predictions/TBBurden2018.csv*).
7. **Results**
   1. **Bac+ prevalence in all districts in Pakistan in 2018**

Absolute TB burden estimates are provided in “08 R script/predictions/AbsoluteTBBurden2018.csv”. The prevalence rate is multiplied by population estimates obtained from district notification data to obtain absolute estimates of TB burden, with districts ranked from highest to lowest. The top 3 districts by absolute burden were Karachi, Rahim Yar Khan and Sialkot, and the lowest 3 were Bannu, Astore and Hunza. In terms of relative burden, the top 3 districts were Kachhi, Mansehra and Sujawal, and the bottom 3 were Larkana, Khuzdar and Kech. See Supplementary Table 1: Absolute district-wise TB burden estimates (*10\_Submission/AbsoluteTBBurden2018.xlsx*).

*Table 2 Top 10 districts with the highest Bac+ prevalence*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| District | province | Estimated Bac Pos Prev | Estimated Bac Pos Prev  (lower 95%CI) | Estimated Bac Pos Prev  (high 95%CI) |
| Karachi | Sindh | 31,420 | 24,883 | 37,697 |
| Rahim yar khan | Punjab | 26,130 | 11,591 | 44,262 |
| Sialkot | Punjab | 22,577 | 17,288 | 27,802 |
| Bahawalnagar | Punjab | 17,797 | 7,895 | 30,146 |
| Mansehra | KP | 15,033 | 12,726 | 17,338 |
| Gujranwala | Punjab | 13,960 | 1,109 | 48,950 |
| Muzaffargarh | Punjab | 12,780 | 286 | 35,853 |
| Bhakkar | Punjab | 12,106 | 10,248 | 13,962 |
| Lahore | Punjab | 11,242 | 7,925 | 14,604 |
| Sanghar | Sindh | 11,194 | 3,801 | 24,518 |

*Table 3 The 10 districts with the lowest Bac+ prevalence*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| District | province | Estimated Bac Pos Prev | Estimated Bac Pos Prev  (lower 95%CI) | Estimated Bac Pos Prev  (high 95%CI) |
| Hunza | GB | 117 | 9 | 410 |
| Astore | GB | 162 | 50 | 211 |
| FR bannu | KP | 251 | 20 | 797 |
| Kech | Balochistan | 267 | 12 | 516 |
| Khuzdar | Balochistan | 270 | 12 | 522 |
| Ghizer | GB | 330 | 26 | 1,157 |
| Sherani | Balochistan | 381 | 30 | 1,337 |
| Ziarat | Balochistan | 382 | 30 | 1,339 |
| Kohlu | Balochistan | 425 | 34 | 1,489 |
| Barkhan | Balochistan | 440 | 35 | 1,544 |

*Table 4 Top 10 districts with the highest Bac+ prevalence rate (per 100,000 population)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| District | province | Estimated Bac Pos Prev Rate  (per 100,000 population) | Estimated Bac Pos Prev Rate  (lower 95%CI) | Estimated Bac Pos Prev Rate  (higher 95%CI) |
| Mansehra | KP | 906 | 767 | 1,045 |
| Bhakkar | Punjab | 906 | 767 | 1,045 |
| Upper dir | KP | 906 | 767 | 1,045 |
| Sujawal | Sindh | 906 | 767 | 1,045 |
| Lakki marwat | KP | 906 | 767 | 1,045 |
| Kachhi | Balochistan | 906 | 767 | 1,045 |
| Tank | KP | 906 | 767 | 1,045 |
| Jhal magsi | Balochistan | 906 | 767 | 1,045 |
| Sialkot | Punjab | 610 | 467 | 751 |
| Haripur | KP | 610 | 467 | 751 |

*Table 5 The 10 districts with the lowest Bac+ prevalence rate (per 100,000 population)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| District | province | Estimated Bac Pos Prev Rate  (per 100,000 population) | Estimated Bac Pos Prev Rate  (lower 95%CI) | Estimated Bac Pos Prev Rate  (higher 95%CI) |
| Kech | Balochistan | 44 | 2 | 85 |
| Khuzdar | Balochistan | 44 | 2 | 85 |
| Larkana | Sindh | 86 | 4 | 251 |
| Faisalabad | Punjab | 86 | 4 | 251 |
| Sukkar | Sindh | 126 | 89 | 164 |
| Lahore | Punjab | 126 | 89 | 164 |
| Tando allah yar | Sindh | 165 | 91 | 251 |
| Matiari | Sindh | 165 | 91 | 449 |
| Nowshera | KP | 165 | 91 | 449 |
| Abbotabad | KP | 165 | 91 | 449 |

* 1. **Predictive power - leave one out cross-validation.**

Statistical measures were applied to Bac+ prevalence estimates per 100 000 population. Using leave-one-out cross-validation, separate Bayesian models for each left-out district were generated and predictions were generated for the district. The output from the district-wise models were compared to the actual BPrev values to calculate the global **RMSE** and **MSE** statistics over all districts. The lowest error scores were obtained using all available data for each district as input to the models, using a two-step process to predict **DistrictHDI** prior to estimating **BPrev**.

*Table 6 MSE*

|  |  |  |
| --- | --- | --- |
| **Method** | **RMSE** | **MSE** |
| All data available per district. DistrictHDI and BPrev Model. | 237.662 | 56483.061 |

**Reference**

1 Alba, S. *et al.* Development and validation of a predictive ecological model for TB prevalence. *International journal of epidemiology* **47**, 1645-1657 (2018).

2 WHO. Tuberculosis prevalence surveys: a handbook. (2011).

3 Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. (Elsevier, 2014).

4 Popescul, A., Pennock, D. M. & Lawrence, S. in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence.* 437-444 (Morgan Kaufmann Publishers Inc.).

5 Russell, S., Binder, J., Koller, D. & Kanazawa, K. *Local learning in probabilistic networks with hidden variables.* in *IJCAI.* 1146-1152.

6 Potgieter, A. & Bishop, J. in *Proceedings of the 2001 International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC.*

7 Guo, C. *et al.* Spatiotemporal analysis of tuberculosis incidence and its associated factors in mainland China. *Epidemiology & Infection* **145**, 2510-2519 (2017).

8 Haase, I. *et al.* Use of geographic and genotyping tools to characterise tuberculosis transmission in Montreal. *The International Journal of Tuberculosis and Lung Disease* **11**, 632-638 (2007).

9 Li, Q. *et al.* The spatio-temporal analysis of the incidence of tuberculosis and the associated factors in mainland China, 2009-2015. *Infection, Genetics and Evolution* **75**, 103949 (2019).

10 Li, X.-X. *et al.* Seasonal variations in notification of active tuberculosis cases in China, 2005–2012. *PloS one* **8**, e68102 (2013).

11 Xiao, Y. *et al.* The influence of meteorological factors on tuberculosis incidence in Southwest China from 2006 to 2015. *Scientific reports* **8** (2018).

12 Yates, T. A. *et al.* The transmission of Mycobacterium tuberculosis in high burden settings. *The Lancet infectious diseases* **16**, 227-238 (2016).

13 Glaziou, P., Sismanidis, C., Zignol, M. & Floyd, K. Methods used by WHO to estimate the global burden of TB disease. *WHO: Geneva, Switzerland* (2016).

14 The Humanitarian Data Exchange. *WorldPop-Pakistan population*, <<https://data.humdata.org/dataset/worldpop-pakistan-population>> (

15 Global Data Lab. *Subnational Human Development Index (SD-2019)*, <<https://globaldatalab.org/shdi/shdi/PAK/>> (

16 Najam, A. & Bari, F. (Islamabad: UNDP, 2017).

17 Verma, A. *et al.* Accuracy of prospective space–time surveillance in detecting tuberculosis transmission. *Spatial and spatio-temporal epidemiology* **8**, 47-54 (2014).

18 Keshavjee, S., Dowdy, D. & Swaminathan, S. Stopping the body count: a comprehensive approach to move towards zero tuberculosis deaths. *The Lancet* **386**, e46-e47 (2015).

19 Theron, G. *et al.* Data for action: collection and use of local data to end tuberculosis. *The Lancet* **386**, 2324-2333 (2015).

20 Gelaw, Y. A., Yu, W., Magalhães, R. J. S., Assefa, Y. & Williams, G. Effect of temperature and altitude difference on tuberculosis notification: A systematic review. *Journal of Global Infectious Diseases* **11**, 63 (2019).

21 Kohonen, T. The self-organizing map. *Proceedings of the IEEE* **78**, 1464-1480 (1990).

22 Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological cybernetics* **43**, 59-69 (1982).

23 Osunmakinde, I. O. & Potgieter, A. in *Proceedings of Southern African Telecommunications Networks and Applications Conference.* 8-9.

24 Osunmakinde, I. O. & Potgieter, A. in *Proceedings of the 10th Southern African Telecommunications Networks and Applications International Conference (SATNAC), Mauritius.*

25 World Intellectual Property Organization. *Patent: (US20040158815) Complex adaptive systems*, <<https://patentscope.wipo.int/search/en/detail.jsf?docId=US40685761&tab=PCTDESCRIPTION&_cid=P12-K028LG-72635-1>>

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)