Subnational estimation of TB prevalence in Pakistan for 2018

Stewart Chang[1]*, William Trouleau[1,2]*, Bradley Wagner[1], Amjad Khan[3], and Jens Levy[4]

1 Institute for Disease Modeling, Bellevue, Washington, USA
2 EPFL Lausanne, Lausanne, Switzerland
3 King Edward Medical University, Lahore, Pakistan
4 KNCV Tuberculosis Foundation, The Hague, Netherlands
* These authors contributed equally

Introduction

Subnational TB prevalence estimates are useful for national TB programs as they create strategic plans, allocate budget and resources for different activities, and coordinate activities with lower levels of government. Working with the Pakistan National Tuberculosis Program, the KIT Royal Institute hosted the 2019 TB Hackathon with the goal of producing "district-wise TB prevalence estimates for Pakistan for 2018 (bacteriologically positive among adults >=15 years) based on the 2010 cluster-level TB prevalence data and other NTP data shared by the Pakistan NTP". The primary challenge was to take a single time-point cross-sectional survey and attempt to estimate subnational TB prevalence at a future time-point. To this end, our team used a logistic model framework, included additional ecological covariates, and exhaustively tested combinations of covariates for the best model fit. We then applied the best-fitting model to generate a map of subnational TB prevalence for Pakistan in 2018.

Methods

Our overall methodology was: (1) to process the prevalence survey microdata and other official data to achieve definitional consistency with published reports and reproduce published findings, (2) to collect and process data from other sources that might serve as ecological covariates, (3) to formulate different models based on a logistic model framework and compare their performance by multiple criteria, and (4) to apply the best-performing model to generating subnational estimates.

Data sources

Official data. Hackathon organizers provided individual-level microdata from the 2010 prevalence survey, shapefiles and population files at the district (admin 2) level, TB case notifications for 2009-2018 at the district level, and other data for Hackathon teams to use, which we refer to as official data.

Additional data. We also searched publicly available sources for additional data including shapefiles and population data at a more granular (subdistrict) resolution than those available in the official Hackathon data. We also searched for ecological covariate data, prioritizing data that were (i) available at granular levels from cross-sectional surveys or as smoothed raster file outputs, (ii) available for both the prevalence survey year (2010-2011) and the target year (2018), and (iii) for factors previously associated with TB disease risk. These factors could be at either the individual level (such as HIV, smoking, diabetes, alcohol use, and malnutrition) or environmental level (such as indoor air pollution and housing density) [1,2]. Below we describe the additional data and reasons for their inclusion as input data:

- *Shapefiles* at tehsil (admin 3) and union council (admin 4) levels from Alhasan Systems Private Limited via the Humanitarian Data Exchange [3]
- *Population estimates* at a 100m level for 2010 and 2018 for total population [4,5] and by age and sex [6,7] from WorldPop
- *Settlement density* at a 100m level for 2010 and 2018 from WorldPop [8]. We were interested to use settlement density as a proxy for urban and rural settings. Urban settings were associated with lower TB prevalence in the published report but were not operationally defined [9].
- *Settlement type* at a 1km level for 2000 and 2015 from the European Commission Global Human Settlement project [10]. Housing types have previously been associated with TB transmission in urban Pakistan as well as in other urban and rural settings [11–13].
- *Multidimensional poverty* at a 1km level for 2007 from WorldPop [14]. Poverty has been associated with TB disease in South Asia, though possibly through other factors such as nutrition [15].
- *Household cooking fuel type, wealth quintile, and urban/rural* designation at a survey cluster level from 2006 and 2017 Demographic and Health Surveys (DHS) [16,17]. Solid cooking fuel has previously been associated with TB transmission in South Asia [18–20], though the evidence from other countries has been mixed [21,22].
- *Individual (female) underweight status and health visit frequency* from 2012 and 2017 DHS surveys [17,23]. Nutrition has a strong association with individual TB disease progression, particularly for BMI ≤ 18.5 [24,25]. We were also interested in health visit frequency as a proxy for overall health as well as access to TB diagnostic services.

Data processing

Prevalence survey. To ensure that our processing pipeline would provide results consistent with published reports [9,26], we attempted to reproduce several indicators from the publications. Overall, we obtained results that were similar to the reported results; however, we identified discrepancies that we could not account for that may also have affected our downstream analysis. All data processing steps were performed in custom Python and R scripts and available in our file repository [27].

- We extracted a slightly different number of **"definite TB cases"** from the microdata compared to the published reports. We focused on this indicator, as the published reports used this indicator to derive overall prevalence and prevalence by cluster and risk group. We implemented the published definition for definite TB cases [9] and extracted 316 cases from the microdata. By comparison, the two published reports stated the survey contained 315 definite cases. A Python file with our extraction pipeline is provided in our file repository [28].

- We obtained a different **cluster-level prevalence distribution** from the published reports. Using definite TB cases, eligible individuals, and cluster IDs from the microdata, we calculated and plotted the raw prevalence rate per cluster. While our distribution had a similar shape to the published distribution (Figure 15 in [9]), our cluster-level rates were smaller than the published values (Figure 1). For example, the survey was reported to contain "one cluster (56) with 14 definite TB cases, leading to a prevalence of definite TB of above 2000 per 100,000" [9]. In the microdata, this cluster contained 14 cases out of 1148 eligible participants resulting in a raw prevalence rate of 1220 per 100 000. Prevalence rates for other clusters were not available in the published reports or from the Hackathon. We hypothesized the published rates may have included imputed cases, as the mean of the binned rates in the published figure (410 per 100 000) was similar to the country-level rate with imputation (398 per 100 000) [26]. However, we did not attempt to reproduce the imputation workflow in the publication and were unable to confirm that this was the source of the discrepancy. A Python file with our cluster-level prevalence derivation is provided in our file repository [29].

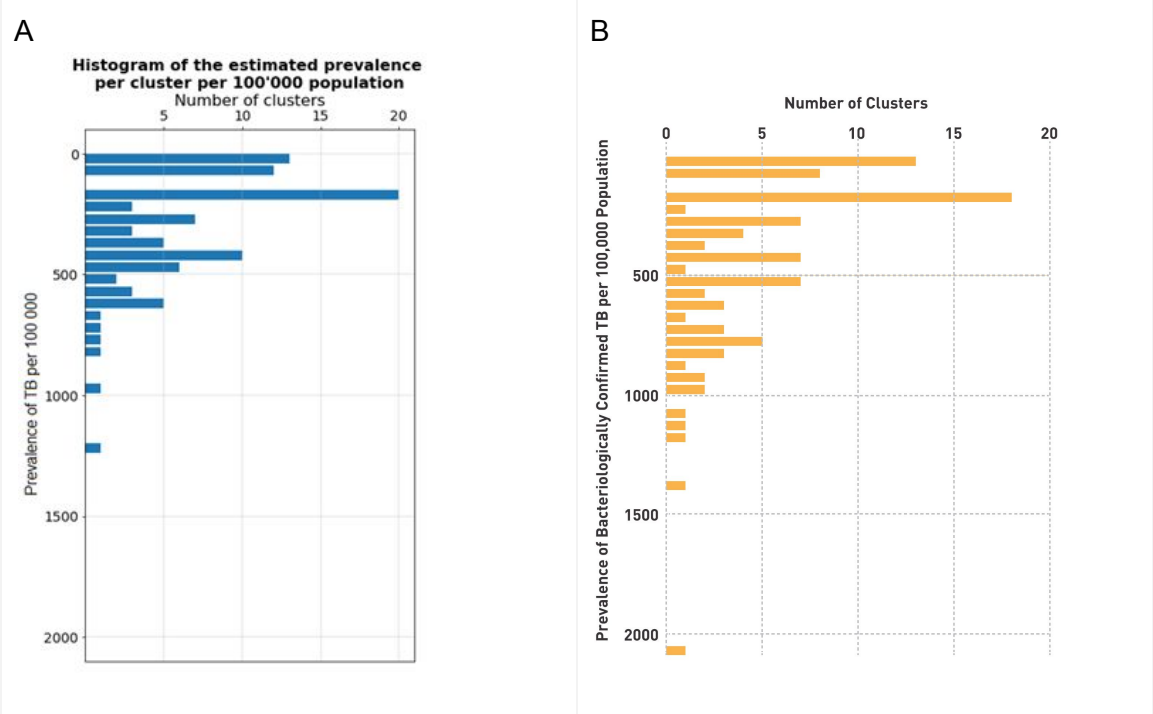


Figure 1. Comparison of rederived cluster-level prevalence rates from the microdata and published rates. (A) As derived from definite TB cases in the microdata. (B) As published (Figure 15 in [9]).

- We also obtained different **prevalence rates for certain risk factor groups** compared to the published reports (Table S1). After applying a logistic model to the individual-level microdata for one factor at a time, we obtained similar results (<2% relative difference) for country-level, sex-specific, age-specific, and urban/rural prevalence rates, with the exception of age 55-64 bin, which had a 5% lower prevalence than reported (Table S1). We also obtained similar prevalence rates for most provinces in Pakistan, e.g., for Punjab and Sindh with <1% relative difference. However, we obtained rates for AJK and Balochistan provinces were 9% lower and 92% higher than reported, respectively, and were unable to account for these differences. A R script with our risk factor-specific prevalence derivation is provided in our file repository [30].

Shapefiles. To map ecological covariates to the prevalence survey clusters, we attempted to obtain the most granular location data possible for the survey clusters. GPS coordinates of the clusters were not made available, so we mapped clusters to the lowest administrative level for which shapefiles were available. This was found to be at the union council (admin 4) level, except in the case of Azad Kashmir and Gilgit-Baltistan, where only tehsil (admin 3) level shapefiles were available. Union council shapefiles were obtained from Alhasan Systems Private Limited via the Humanitarian Data Exchange [3], and union council names were matched to survey cluster union council names using automatic fuzzy name matching available in the SpatioTemporal Analysis and Mapping in Python (STAMP) tool from the Institute for Disease Modeling [31]. Remaining cluster union council names were matched manually through Internet searches for alternative spellings or historical name changes. Cluster-level prevalence rates were mapped to their respective admin locations and plotted (Figure 1A). A Python file with our shapefile processing pipeline is available in our file repository [32].

Simple covariate raster extraction. Several covariates of interest including multidimensional poverty, settlement density, and settlement types were already publicly available as high resolution population raster files (Table 1). In these cases, we overlaid the union council shapefile, extracted raster pixels within the union council boundaries, and derived a summary value for each union council, e.g., the proportion of pixels positive for built-up settlement (Table 1). When we compared these values to cluster-level prevalence rates, we found that values for multidimensional poverty, settlement density, and settlement types were all found to be correlated with cluster-level TB prevalence (Table 1). R scripts for the extraction and comparison of covariate values are available in our file repository [33].

DHS covariate extraction. Special consideration was made for DHS survey data, as DHS surveys contain multiple indicators relevant to TB, were conducted at a high density throughout the country, and were available for several years (DHS characteristics in Table S2). Household-level indicators including solid cooking fuel usage, wealth quintile status, and

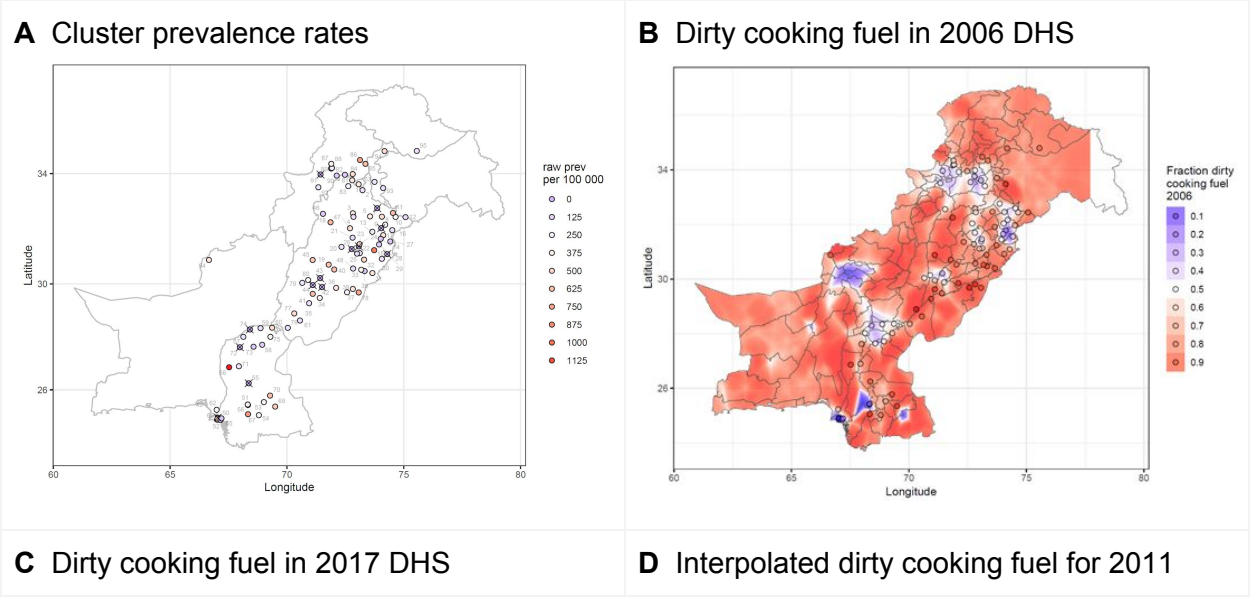
urban/rural classification were measured with GPS coordinate availability in DHS 2006 and 2017 (Table S2). In these cases, we downloaded the DHS microdata and derived a summary value for each DHS survey cluster, e.g., the proportion of households using a solid cooking fuel type (Table 1). Based on these values, we generated a smoothed (kriged) surface for the entire country using mapping software available from the Institute for Disease Modeling (Figure 2B-C, [34]). Values were extracted for union councils that contained TB prevalence survey clusters, and from the summary values for 2006 and 2017, we derived values for 2011 by linear interpolation (Figure 2D). Similarly, we extracted individual-level indicators for the proportion of adult females who were underweight or had not visited a health clinic recently from DHS 2012 and 2017 (Table S2), generated a smoothed surface of summary values for the entire country, and extracted values at the union council level (Table 1). Out of the DHS indicators that we considered, cooking fuel type, wealth quintile status, urban/rural status, and underweight status were all found to be correlated with cluster-level TB prevalence (Figure 2E, Table 1).

Pakistan TB routine data extraction. TB notifications by age and sex at the district level for multiple recent years were provided by the Hackathon. We calculated the bacteriologically positive and total case notification rates (CNR) for each district in 2011 and compared these to TB prevalence rates for each cluster. However, we did not observe a significant correlation between either of these indicators and cluster-level TB prevalence rate and did not carry them forward into the model selection step (Figure 1F, Table 1). We hypothesized that accounting for the level of TB diagnostic access at the time of data collection would increase the association with TB prevalence; however, data on the number of TB clinics by district, travel time to TB clinic, or other proxies for diagnostic access for 2011 were unavailable. Likewise, HIV notifications by province were provided by the Hackathon, but we did not include HIV indicators in our model selection step due to the coarse granularity of these data and a lack of additional data on HIV testing access. R scripts for these analyses are available in our file repository [33].

Name (level)	**Source (type)**	**Value derived at union council level**	**2018 available?**	**Corr with raw prev rate (p val)**
Solid cooking fuel (cluster)	DHS 2006, 2017 (survey)	Proportion of households using solid cooking fuel	Yes (2017)	0.31 (0.002)
Lowest wealth quintile (cluster)	DHS 2006, 2017 (survey)	Proportion of households in lowest quintile	Yes (2017)	0.28 (0.005)
Rural by DHS classification (cluster)	DHS 2006, 2017 (survey)	Proportion of households in rural setting	Yes (2017)	0.28 (0.006)
Underweight (cluster)	DHS 2012, 2017 (survey)	Proportion of women with BMI < 18.5	Yes (2017)	0.18 (0.07)
Multidimensional poverty index, MPI (1km)	WorldPop 2007 (modeled)	Proportion of people in poverty	No	0.17 (0.11)

Built settlement growth model, BSGM (100m)	WorldPop 2000-2015 (modeled)	Proportion of land with built-up settlement	Yes (extrapolated)	-0.16 (0.13)
				-0.19 (0.07): asin sqrt
GHS Settlement Model type, GHS-SMOD (1km)	European Commission 2000, 2015 (modeled)	Proportion of land with most rural settlement type	No	0.16 (0.13)
				0.17 (0.10): asin sqrt
Bacteriologically positive case notification rate (district)	Pakistan NTP 2009-2018 (routine data)	Bacteriologically positive cases per capita per year	Yes (2018)	0.10 (0.32)
No health visit in last 12 mo (cluster)	DHS 2012, 2017 (survey)	Proportion of women without health visit in last 12 mo	Yes (2017)	0.07 (0.49)
Total case notification rate (district)	Pakistan NTP 2009-2018 (routine data)	Total (bact positive, bact negative, and extrapulm) cases per capita per year	Yes (2018)	-0.02 (0.88)

Table 1. Additional health-related risk factor covariates considered for inclusion in the model. Covariates were extracted at the union council level and compared to cluster-level prevalence rates. Covariates having correlation coefficient $p < 0.2$ were considered for use in the subsequent model selection step. Rows ordered by increasing correlation coefficient p-value.



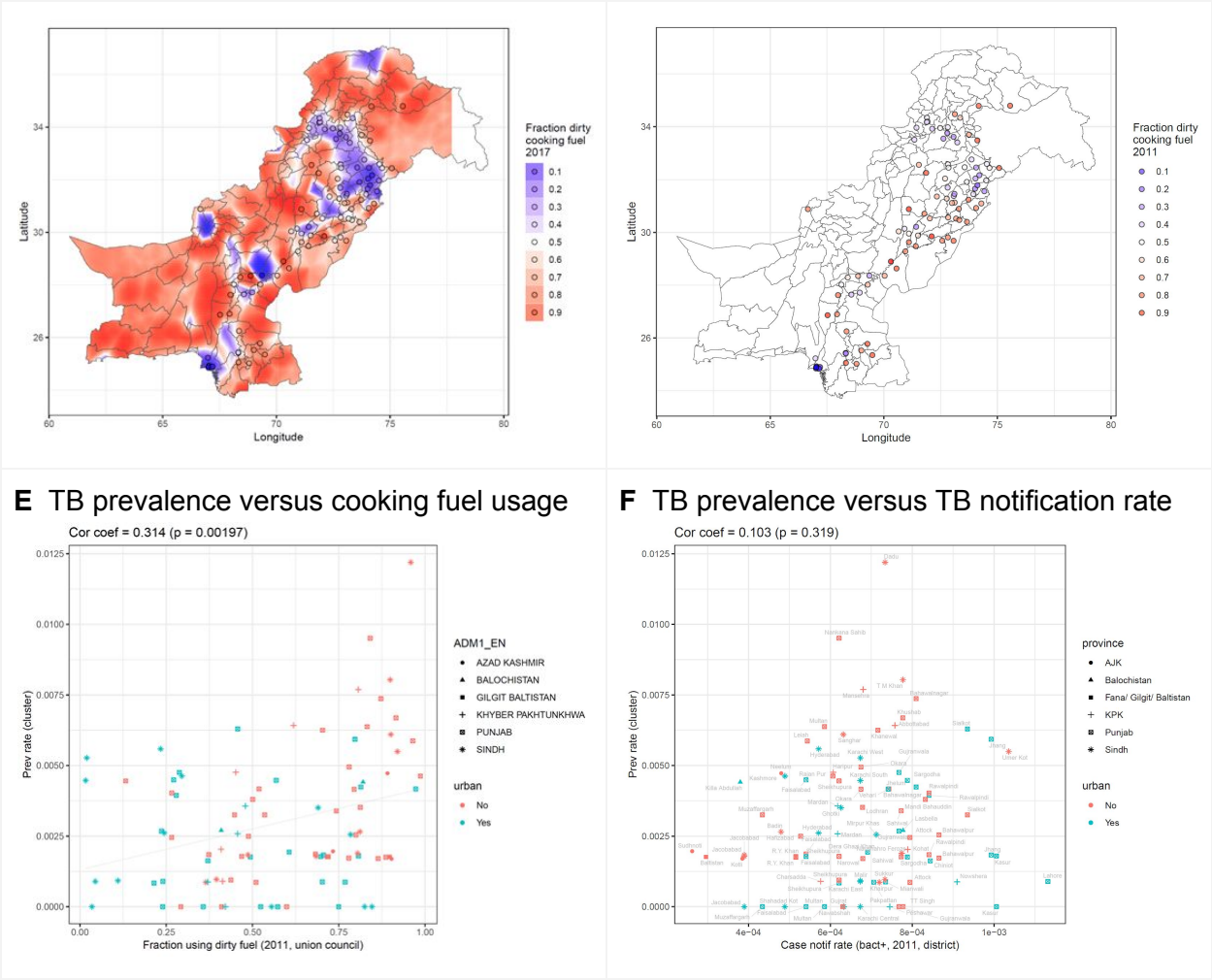


Figure 2. Illustration of covariate testing for inclusion in the predictive model. (A) Cluster-level TB prevalence rates mapped to the union council level. "X" indicates clusters with zero TB cases in the survey. (B-C) Dirty (solid) cooking fuel usage as smoothed (kriged) from DHS 2006 and 2017 survey data. Circles represent TB prevalence survey union councils for reference. (D) Dirty cooking fuel usage extracted for year 2011 at the union council level. (E) Comparison of dirty cooking fuel usage at the union council level to cluster-level TB prevalence. (F) Comparison of TB case notification rates at the district level to cluster-level TB prevalence.

Model formulation

We used variations of the binomial-logistic model to represent TB prevalence. Logistic models are well suited for survey data, where the number of individuals testing positive in a cluster can be considered a binomial random variable $(X_{cluster})$ given the total number of individuals in the cluster $(n_{cluster})$ and a probability $(\pi_{cluster})$ determined by cluster-level predictors $(x_{i,\,cluster})$ with a logit link function. Alternatively, such a model can also be applied at the individual level, where $n_{indiv} = 1$ replaces $n_{cluster}$ and additional individual-level predictors $(x_{i,\,individual})$ are considered. We also adopted a logistic model framework to achieve consistency with WHO guidelines, as

guidelines for the design and analysis of TB prevalence surveys recommends the use of logistic models [35]. Our basic formula was:

$$X_{cluster} \sim Binomial(n_{cluster}, \pi_{cluster})$$

$$logit(\pi_{cluster}) = b_0 + \sum_i \sum_{j \neq i} x_{i,\,cluster}(\beta_i + \beta_{i:j}\, x_{j,\,cluster}) + u_{cluster}$$

where b_0 is a country-level intercept, $x_{i,\,cluster}$ is a cluster-level covariate that could also include individual-level covariates aggregated at the cluster level, $x_{j,\,cluster}$ is another cluster-level covariate that interacts with $x_{i,\,cluster}$, β_i and $\beta_{i:j}$ are coefficients for covariates and covariate interaction terms, respectively, and $u_{cluster}$ is an optional cluster-level random effect. In addition, we applied a Bayesian approach to model fitting which allowed us to infer parameter distributions that may have been non-Gaussian and fit hyperparameters shared between clusters such as the variance for cluster-level intercepts. Models were formulated in the R statistical language using the base glm function which allowed fixed effects, the lme4 package function glmer which allowed for random effects including random intercepts [36], and the R-INLA package function inla which allowed for Bayesian sampling of posterior distributions for both fixed and random effects [37].

The covariates that we considered for model predictors included demographic, geographic, and additional health-related risk factor covariates (Table 2). Demographic and geographic covariates were obtained directly from the prevalence survey individual-level data and aggregated at the cluster level. Additional health-related risk factor covariates included those correlated with cluster-level TB prevalence and were aggregated at the union council level (Table 1). We observed evidence of multicollinearity and possible interactions when we examined the covariate data in total. Some covariates appeared to be multicollinear with others, particularly those related to poverty, which suggests that including multiple poverty-related indicators may not necessarily improve model fit (Figure S1). We also observed evidence of interactions, particularly with geographic (province) factors (Figure S2). For example, clusters with a high proportion of older adults (aged 55 and over) were found to have higher TB prevalence rates but only in Sindh province, particularly in rural areas (Figure S2). Likewise, a principal components analysis of the data showed that clusters with high proportions of both younger adults (aged 15-24) and female adults were associated with lower TB prevalence rates (Figure S3).

Covariate type	**Variable name**	**Description (in cluster/union council level data)**	**Interaction position (i:j)**
Demographic (from survey or population raster)	age_1524	Proportion of cluster population aged 15-24	i
	age_2534	Proportion of cluster population aged 25-34	i

	age_3544	Proportion of cluster population aged 35-44	i
	age_4554	Proportion of cluster population aged 45-54	i
	age_5564	Proportion of cluster population aged 55-64	i
	age_65up	Proportion of cluster population aged 65+	i
	female	Proportion of cluster population female	i, j
Geographic (from survey or population raster)	punjab	Cluster in Punjab?	i, j
	sindh	Cluster in Sindh?	i, j
	kpk	Cluster in KPK?	i, j
Additional risk factors	dirty_fuel	Proportion of households in u.c. using solid fuel (DHS)	i
	poorest	Proportion of households in u.c. in poorest 20% (DHS)	i
	rural	Proportion of households in u.c. rural (DHS)	i
	underweight	Proportion of adult females in u.c. with BMI ≤ 18.5 (DHS)	i
	asin_bsgmi	Proportion of u.c. with BSGM index = 1 (arcsin sqrt)	i
	bsgmi_urban	Proportion of u.c. with >15% area with BSGM index = 1	j
	asin_settle_type_poor	Proportion of u.c. with SMOD type 11 (arcsin sqrt)	i
	povMPI_popWeighted	Proportion of u.c. in poverty in 2007	i

Table 2. Candidate covariates used during model selection. Interaction position indicates how the covariate was used in constructing interaction terms. Covariates were used individually and as part of linear combinations $C_{n,i}$ where $n = 18,\ i \in \{1, 2, 3, 4\}$. Interaction terms were constructed from a subset of the covariates where each interaction term was indicated by $C_{m,i} \times C_{n,j}$ where $m = 14,\ i = 1$ and $n = 5,\ j = 1$, and the number of interaction terms $k \in \{1, 2, 3\}$. In addition, formulas based on additional combinations of interest were added manually.

Model selection

We specified models based on the logistic regression framework that included different sets of covariates with or without interactions (Table 2). For each model, we used a leave-one-out (LOO) cross-validation (CV) approach where each cluster was held out and data for the remaining clusters were used to train the model and derive its coefficients. The fitted model was used to predict the response (prevalence rate) using the covariate data from the held-out cluster. After iterating through all clusters, we computed mean squared error (MSE), residual

standard error (RSE), and coefficient of determination (R^2). To expedite the selection process, we initially applied only fixed effects models with maximum likelihood estimates to the whole set of models where each model was specified by a different set of covariates. In total we tested 261,635 different models. For the best-scoring models, scoring was redone with fixed and random effects with posterior sampling using the INLA library. Overfitting was assessed by excluding the top 10% highest prevalence clusters (Clusters 11, 15, 18, 40, 41, 56, 67, 78, 85, and 86) and repeating LOO-CV with the best-scoring models on the remaining clusters.

Model prediction

We used the best-fitting model (Table 3) to derive district-level TB estimates for Pakistan in 2018. This model was refit on the entire prevalence survey data set and covariate data set from 2011, resulting in 1000 parameter draws for the set of coefficients. These coefficients were applied to the raster-level data for the entire country for 2018 aggregated at a 5-km resolution. This resulted in 1000 predictions of TB prevalence for each 5-km pixel. These values were weighted by the overall adult population in each pixel and extracted at the district level, resulting in predictions for the population-weighted average prevalence in each district. We used these values to generate district-level summary statistics (mean, standard deviation, and 95% confidence intervals). Predictions for each district are provided in the accompanying CSV file.

Code reproducibility

Our data processing scripts coded in Python and R and our model scripts coded in R are available in our GitHub file repository [27]. In addition to this main repository, which contains various ancillary scripts that were not used in the final prediction, we have created a second repository that hosts an essential subset of scripts for generating model predictions ([38], screenshot in Figure S4).

Results

Model performance

To determine the best-fitting models to the prevalence survey data, we tested different combinations of covariates, individually and in interaction terms, in a logistic model framework with cluster-level LOO-CV. This yielded a set of best-fitting models which had an RSE of approximately 0.7 and R^2 of approximately 0.3 (Table 3, Figure 3). The best-fitting models achieved higher scores primarily by fitting to the highest prevalence clusters better than other models (Figure 3, Table S3). Among the best-fitting models, we noted several similarities. Many contained demographic covariates associated with low prevalence rate clusters, namely the proportion of adults aged 15-24 or female (Table 3). These might be considered protective factors at the population level and were associated with lowest TB prevalence quartile clusters

in the initial PCA (Figure S3). Likewise, many best-fitting models also contained covariates correlated with high prevalence rates, namely the proportion of adults aged 65 and over, particularly in Sindh province (Table 3). These might be considered high risk factors at the population level and were also observed during our initial data analysis (Figure S2). By comparison, prevalence rates in Punjab province were predicted to be relatively homogeneous.

Rank	Model formula	MSE	RSE	R^2
1	response = (age_1524 * female) + (sindh + age_65up : sindh) + (kpk + underweight : kpk)	39210	0.695	0.320
2	response = (age_1524 * female) + (age_65up * sindh) + (kpk + underweight : kpk)	39856	0.706	0.310
3	response = (age_1524 * female) + (sindh + age_65up : sindh) + (underweight * kpk)	40429	0.716	0.303
4	response = (underweight * female) + (age_65up * sindh) + (underweight * kpk)	40968	0.726	0.285
5	response = (age_1524 * female) + (dirty_fuel * female) + (age_5564 * bsgmi_urban)	41244	0.731	0.278
6	response = (age_1524 * female) + (age_65up * sindh) + (underweight * kpk)	41248	0.731	0.292

Table 3. Best-fitting models by cluster-level LOO-CV by MSE, RSE, and R^2. In the formula notation, $x_1 : x_2$ represents factor multiplication, while $x_1 * x_2$ represents factor crossing and is equivalent to $x_1 + x_2 + x_1 : x_2$. A country-level intercept is assumed for all models. Note that MSE was based on predicted and observed values with units of TB cases per 100 000.

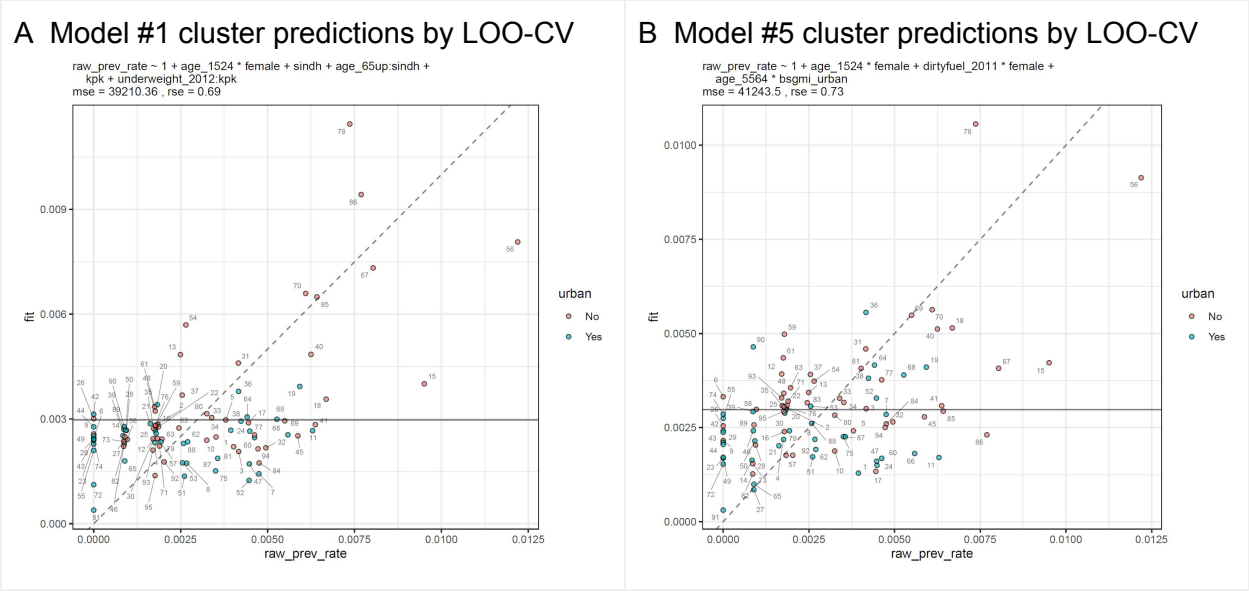


Figure 3. Cluster-level prevalence rates predicted by two models using LOO-CV compared to observed values. (A) Model #1 from Table 2. (B) Model #5 involving a different set of covariates from Table 2. The horizontal line represents the country-level intercept, while the diagonal line represents the x = y line of perfect fit.

Model prediction of subnational prevalence

Our subnational map of TB prevalence, based on our overall best-fitting model (Table 3) applied to the available covariate data for 2018, shows the presence of geographic heterogeneities in TB prevalence across Pakistan. In particular, our map predicts that the highest TB prevalence districts may be found in Sindh province (Figure 4A). Out of the ten highest predicted prevalence districts in Pakistan, eight were located in Sindh province, while the remaining two were located in KPK (see accompanying CSV file). The patterning of high prevalence rates in Sindh, particularly in the southeast area of this province, was consistent with the presence of high-prevalence survey clusters in this area (clusters 67, 69, and 70, which were in the top quintile of prevalence rates, cf. Figure 1A). However, the predicted prevalence rates for high prevalence districts were also accompanied by large uncertainties (Figure 4B). By comparison, districts in Punjab were predicted to be relatively homogeneous with respect to TB prevalence rates (Figure 4A).

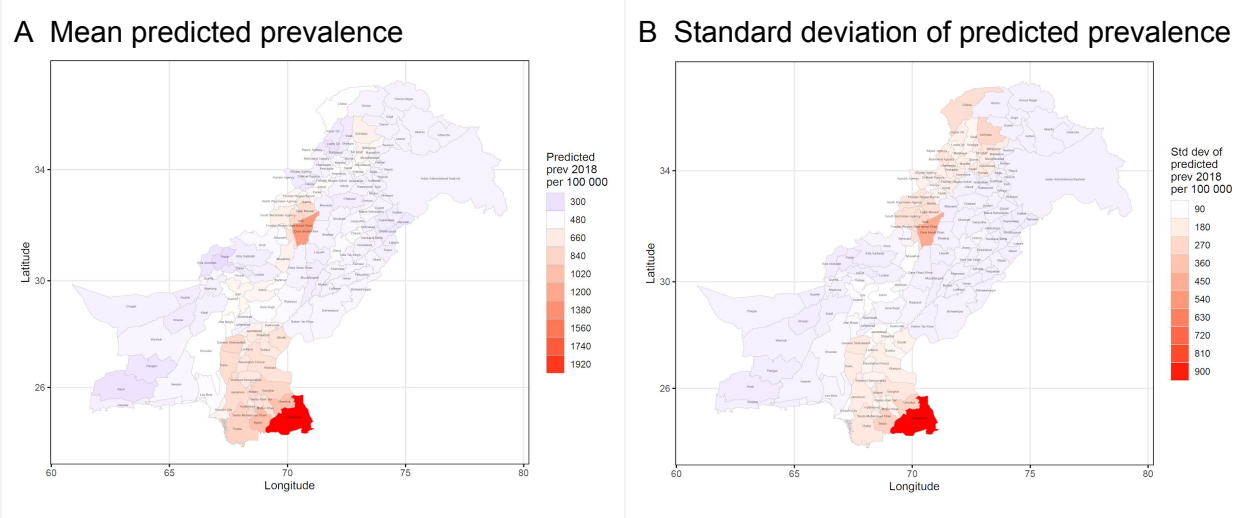


Figure 4. Predicted district-level TB prevalence for 2018. (A) Mean predicted prevalence. (B) Standard deviation of predicted prevalence from repeated draws of posterior distribution.

Discussion

In this report we present one possible approach for estimating subnational TB prevalence when a single prevalence survey is available to serve as a gold standard but was not powered to provide subnational estimates or estimates for particular risk groups. In our approach, we used a Bayesian binomial logistic regression model and included a broad set of covariates, including newly data derived from DHS surveys, that were tested singly and in combination to derive a model of best-fit to the prevalence survey data.

Drivers of TB. Our best-fitting model contained terms that reflected the impact of geography (as province) on other risk factors. For example, while older populations were generally associated with higher TB prevalence, this appeared especially to be the case in Sindh province. Similarly, while populations with more underweight adults tended to have higher TB prevalence, this effect was more pronounced in KPK province. However, in both cases, the effect was informed by a relatively sparse number of clusters and may be susceptible to sampling bias and lead to overfitting. However, if validated in further studies, these results suggest that different case-finding strategies and interventions may be needed in different provinces in Pakistan.

Strengths and limitations. Strengths of our approach included the use of a standard logistic modeling approach that the WHO has recommended for analyzing prevalence survey data [35]. In addition, we used demographic covariates (e.g., binned age) and geographic covariates (e.g., province designations) that were identical to those in the published reports [9,26]. Therefore, the model structure and many of the covariates should be interpretable to national TB program officers and TB researchers. To this framework, we added covariates that have strong connections to health and TB disease; these included undernutrition, indoor air pollution from solid cooking fuel use, and poverty level. We derived many of these as modeled surfaces directly from the DHS surveys. While the DHS program has made modeled raster surfaces available from its surveys for selected indicators, years, and countries [39], no such surfaces were available for Pakistan at the time of the Hackathon. To our knowledge, spatial data in Pakistan for undernutrition as indicated by low BMI, indoor air pollution from solid cooking fuel usage, and access to health clinics were not available from any other source. In particular, undernutrition and indoor air pollution have been associated with approximately 3-fold and 1.4-fold increases in TB risk, respectively [1,2]. We also employed an exhaustive model selection algorithm that allowed combinations of covariates to be quickly assessed for their utility in the model, followed by a final Bayesian-based model fitting procedure that allowed full posterior distributions for model coefficients to be extracted.

Limitations to our approach included the considerable variance that remained unexplained in the cluster-level prevalence rates. For example, our best-fitting model accounted for only about 30% of the variance in the data. In particular, we were unable to identify distinguishing characteristics of clusters with zero observed cases during the survey. One explanation is that none of our covariates, which included known TB risk factors, could account for this type of outcome. In other words, while the presence of a risk factor (such as high undernutrition) may increase the risk of a high prevalence of TB in the community, the absence of the risk factor may not necessarily guarantee that the community will have zero TB cases. Another explanation is that our data were too coarse to identify characteristics associated with low TB clusters. While we extracted data to the lowest matchable administrative level possible, this level (union council) remains larger than the survey clusters, and more geographically heterogeneous risk factors may not be captured in our data. It is also possible that other covariates such as land use and rainfall that lack an established connection to TB disease and were not included in our analysis may actually account for low TB clusters.

Overfitting. Overfitting was a particular concern. We noted that the best-fitting models tended to have a better fit to the highest prevalence clusters without necessarily having a better fit to lower prevalence clusters compared to other models (Figure 3A). To test this hypothesis, we removed high prevalence clusters from the dataset, refit the best-scoring models, and observed that performance by LOO-CV was significantly decreased (Table S3). Therefore, the best-fitting models that we identified may have extracted characteristics particular to the highest prevalence areas in Pakistan. We attempted to improve the fit to lower prevalence clusters using variants of the binomial likelihood that allowed for additional hyperparameters for variance such as the beta-binomial or zero-inflated-binomial model [40], but we did not observe improvements in performance by LOO-CV (data not shown).

Future work. While we originally intended to use TB case notification data in our model, we were unable to find a transformation of the routine notification data to make it predictive of TB prevalence. We hypothesize that additional granular health access data such as the density of TB diagnostic centers at the time of the survey may allow us to connect notifications to prevalence. Further collaboration with local TB officials and the national TB program may allow these historical data to be recovered and subsequently a covariate to be added to the model that predicts the impact of programmatic activity on TB prevalence.

Appendix

Characteristic	**Levels**	**Point prevalence estimate (microdata reanalysis)**	**Point prevalence estimate (as reported in [9])**	**Relative difference in mean estimates**
Definite TB	NA	298.4 (267.2, 333.1)	296.6 (248.1, 345.2)	+0.61%
Sex	Male (ref level)	366.0 (314.1, 426.4)	365.1 (293.7, 436.6)	+0.25%
	Female	248.8 (190.2, 325.3)	246.9 (196.2, 297.6)	+0.77%
Age Group	15-24 y (ref level)	172.1 (133.1, 222.5)	174.8 (124.7, 225)	-1.54%
	25-34 y	153.9 (95.2, 248.8)	154.6 (99.1, 210.2)	-0.45%
	35-44 y	284.1 (181.6, 444.4)	287.4 (194.6, 380.1)	-1.15%
	45-54 y	385.5 (244.2, 608.1)	393.8 (272.7, 514.8)	-2.11%
	55-64 y	**457.7 (279.2, 749.6)**	**479.6 (303.6, 655.5)**	**-4.57%**
	65+ y	1167.9 (765.5, 1778.1)	1170.2 (870.8, 1469.5)	-0.20%
Type of Area	Rural (ref level)	354.5 (309.3, 406.3)	351.8 (280.9, 422.7)	+0.77%
	Urban	230.0 (175.9, 300.9)	229.9 (171.6, 288.2)	+0.04%
Province	Punjab (ref level)	289.7 (251.0, 334.4)	291.8 (234.4, 349.2)	-0.72%

	Sindh	311.8 (230.9, 421.1)	313.2 (190.6, 435.8)	-0.45%
	Balochistan	**356.7 (173.4, 732.8)**	**185.8 (170.7, 200.9)**	**+91.98%**
	AJK	**276.9 (139.8, 548.1)**	**304.3 (159.6, 449)**	**-9.00%**
	KPK	327.5 (220.3, 486.7)	320.2 (154.6, 485.9)	+2.28%
	Gilgit-Baltistan	176.5 (43.6, 712.4)	176.5 (NA, NA)	+0.00%

Table S1. Comparison of prevalence values from our reanalysis of the microdata versus published values [9]. We applied a logistic model with one fixed effect per characteristic (factor) at a time with the stated levels and without additional cluster random effects to the individual-level data in the R statistical software. Confidence intervals were computed on the basis of the sum of the variance of the reference level coefficient and stated level coefficient. Note that the inverse logit function applied to the resulting normally distributed confidence intervals yields asymmetric confidence intervals compared to the symmetric confidence intervals in the published report. Deviations greater than 3% are indicated in bold.

Indicator (type, variable)	**2006-07**	**2012-13**	**2017-18**
GPS coord or district only?	GPS	District only (map to district centroid)	GPS
BMI among women (continuous, ha40/v445)	No	Yes	Yes
Cigarettes smoked in last 24h (discrete, ha35/mv464)	No	Yes	Yes
Visited health facility in last 12mo (yes/no, v394)	No	Yes	Yes
Cooking fuel (categorical, hv226)	Yes	Yes	Yes
Residence place type (urban/rural, hv025)	Yes	Yes	Yes
Residence place (categorical, hv026)	Yes	Yes	No
Relative wealth index (categorical, hv270)	Yes	Yes	Yes
Relative wealth score (continuous, hv271)	Yes	Yes	Yes

Table S2. Availability of TB-relevant indicators in three recent Pakistan DHS surveys.

Orig rank	**Model formula**	**MSE (all)**	**MSE (90%)**	**R2 (all)**	**R^2 (90%)**
1	response = (age_1524 * female) + (sindh + age_65up : sindh) + (kpk + underweight : kpk)	0.695	1.059	0.320	0.010
2	response = (age_1524 * female) + (age_65up * sindh) + (kpk + underweight : kpk)	0.706	1.082	0.310	0.004
3	response = (age_1524 * female) + (sindh + age_65up : sindh) + (underweight * kpk)	0.716	1.087	0.303	0.005

4	response = (underweight * female) + (age_65up * sindh) + (underweight * kpk)	0.726	1.165	0.285	0.001
5	response = (age_1524 * female) + (dirty_fuel * female) + (age_5564 * bsgmi_urban)	0.731	0.925	0.278	0.109
6	response = (age_1524 * female) + (age_65up * sindh) + (underweight * kpk)	0.731	1.113	0.292	0.001

Table S3. Assessment of overfitting by exclusion of clusters with the highest 10% of prevalence rates (Clusters 11, 15, 18, 40, 41, 56, 67, 78, 85, and 86). Each model was assessed by LOO-CV on the remaining 85 clusters. MSE and R^2 results are shown for all clusters versus 85 clusters.

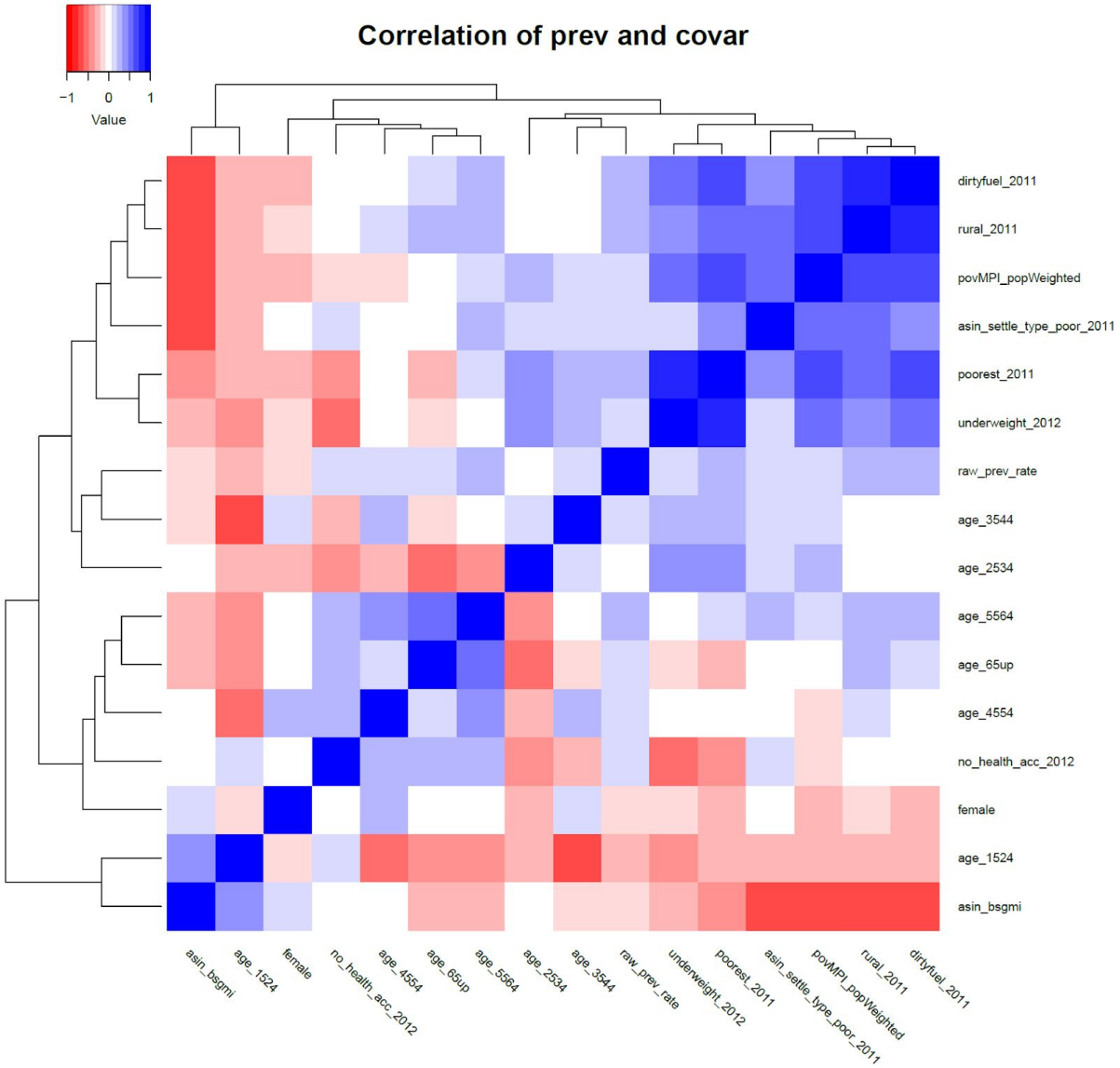


Figure S1. Evidence of multicollinearity in the covariate data. Pairwise correlation coefficients of cluster-level prevalence rates ("raw_prev_rate"), prevalence survey demographic groups (e.g.,

"age_1524") and ecological covariates extracted at the union council level were plotted. Note the strong similarity between the poverty-related covariates.

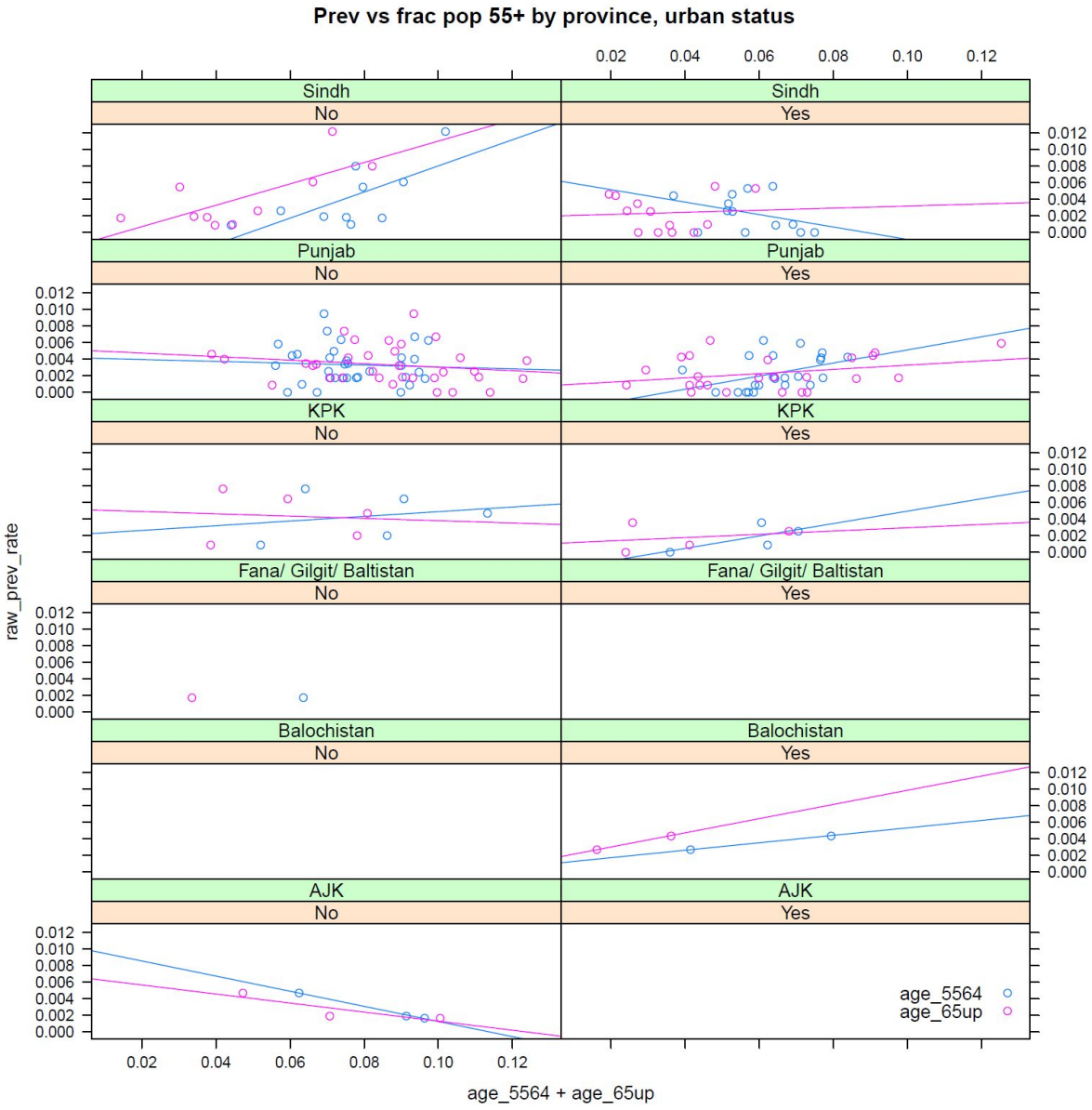


Figure S2. Illustration of interactions in the covariate data. Populations with high numbers of older adults (age ≥55 years) were associated with higher cluster-level TB prevalence rates but only in Sindh and rural areas.

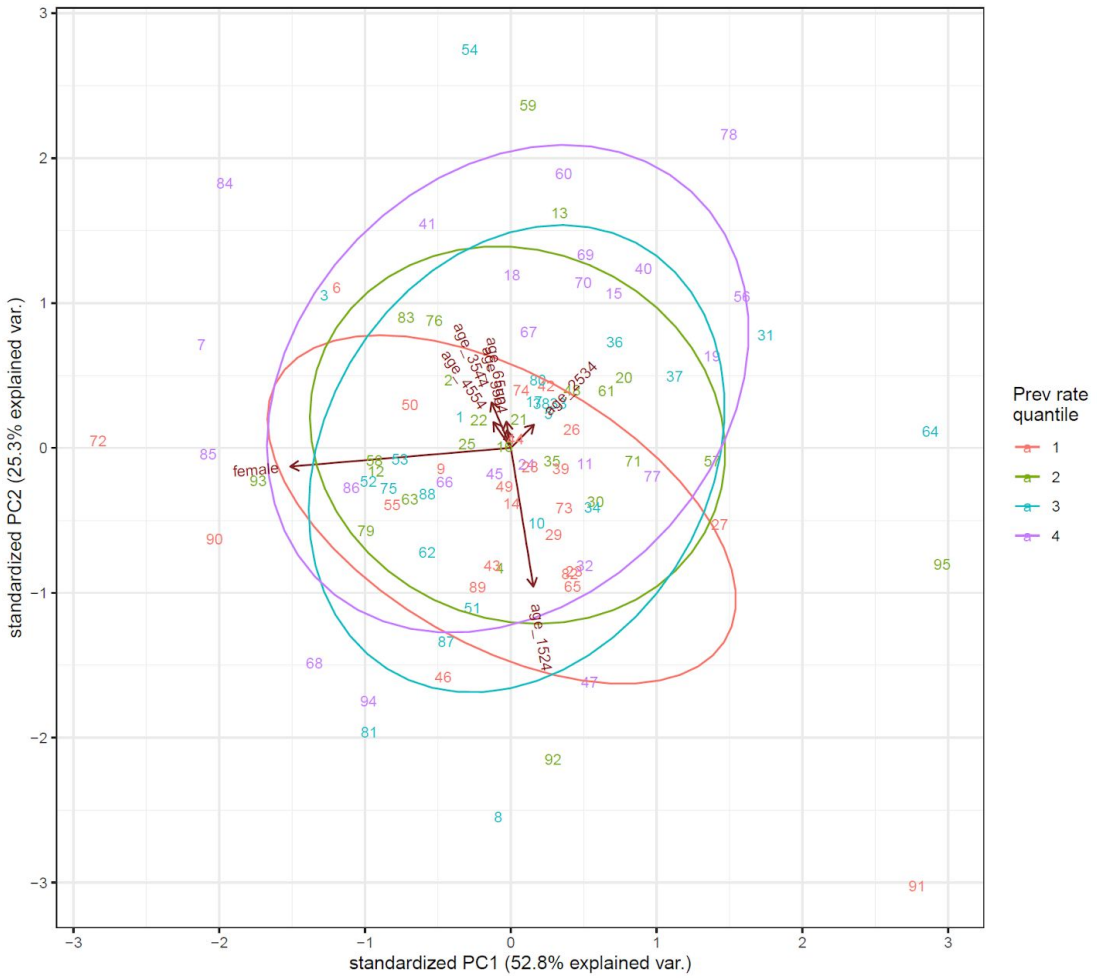


Figure S3. PCA of the prevalence survey demographic data indicates that the proportion of females and adults aged 15-24 tended to characterize clusters in the lowest prevalence rate quantile. Cluster numbers are plotted with colors indicating the lowest (quantile = 1) to highest (quantile = 4) prevalence rates.

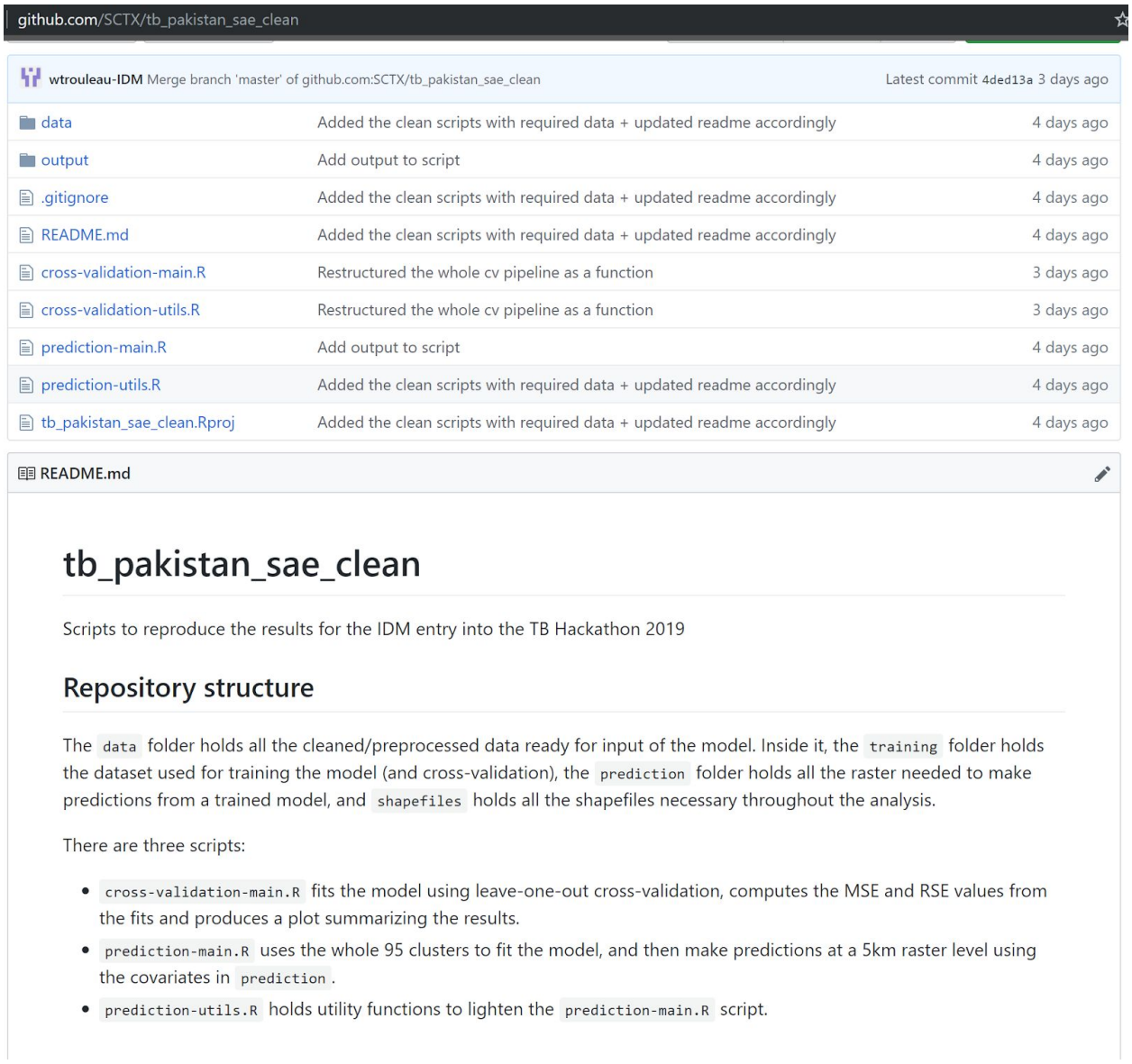


Figure S4. Screenshot of our GitHub file repository with code to reproduce the prediction pipeline at https://github.com/SCTX/tb_pakistan_sae_clean.

References

1. Dye C, Lönnroth K, Jaramillo E, Williams BG, Raviglione M. Trends in tuberculosis incidence and their determinants in 134 countries. Bull World Health Organ. 2009;87: 683–691. Available: https://www.ncbi.nlm.nih.gov/pubmed/19784448

2. Lönnroth K, Castro KG, Chakaya JM, Chauhan LS, Floyd K, Glaziou P, et al. Tuberculosis control and elimination 2010-50: cure, care, and social development. Lancet. 2010;375: 1814–1829. doi:10.1016/S0140-6736(10)60483-7

3. Pakistan Union Council Boundaries along with other admin boundaries dataset [Internet].

Available:
https://data.humdata.org/dataset/pakistan-union-council-boundaries-along-with-other-admin-boundaries-dataset

4. The spatial distribution of population in 2010, Pakistan [Internet]. Available: https://www.worldpop.org/geodata/summary?id=3924

5. The spatial distribution of population in 2018, Pakistan [Internet]. Available: https://www.worldpop.org/geodata/summary?id=5916

6. Pakistan 100m Age structures in 2010 [Internet]. Available: https://www.worldpop.org/geodata/summary?id=14398

7. Pakistan 100m Age structures in 2018 [Internet]. Available: https://www.worldpop.org/geodata/summary?id=16390

8. Built-Settlement Extents, Pakistan [Internet]. Available: https://www.worldpop.org/geodata/summary?id=17217

9. Qadeer E, Fatima R, Tahseen S, Samad Z, Kalisvaart N, Tiemersma E, et al. Prevalence of Pulmonary Tuberculosis among the Adult Population of Pakistan 2010-2011 [Internet]. 2013. Available: http://www.ntp.gov.pk/uploads/Prevalence_Report.pdf

10. GHS-SMOD [Internet]. Available: https://ghsl.jrc.ec.europa.eu/ghs_smod2019.php

11. Haider BA, Akhtar S, Hatcher J. Daily contact with a patient and poor housing affordability as determinants of pulmonary tuberculosis in urban Pakistan. Int J Mycobacteriol. 2013;2: 38–43. doi:10.1016/j.ijmyco.2012.12.003

12. Low C-T, Lai P-C, Tse W-SC, Tsui C-K, Lee H, Hui P-K. Exploring tuberculosis by types of housing development. Soc Sci Med. 2013;87: 77–83. doi:10.1016/j.socscimed.2013.03.024

13. Clark M, Riben P, Nowgesic E. The association of housing density, isolation and tuberculosis in Canadian First Nations communities. Int J Epidemiol. 2002;31: 940–945. doi:10.1093/ije/31.5.940

14. Pakistan 1km Poverty [Internet]. Available: https://www.worldpop.org/geodata/summary?id=1275

15. Oxlade O, Murray M. Tuberculosis and poverty: why are the poor at greater risk in India? PLoS One. 2012;7: e47533. doi:10.1371/journal.pone.0047533

16. Pakistan: Standard DHS, 2006-07 [Internet]. Available: https://dhsprogram.com/what-we-do/survey/survey-display-273.cfm

17. Pakistan: Standard DHS, 2017-18 [Internet]. Available: https://dhsprogram.com/what-we-do/survey/survey-display-523.cfm

18. Magheswari U, Johnson P, Ramaswamy P, Balakrishnan K, Jenny A, Bates M. Exposure to Biomass Fuel Smoke and Tuberculosis—A Case-Control Study in India. Epidemiology.

2007;18: S122. doi:10.1097/01.ede.0000276689.31498.30

19. Sehgal M, Rizwan SA, Krishnan A. Disease burden due to biomass cooking-fuel-related household air pollution among women in India. Glob Health Action. 2014;7: 25326. doi:10.3402/gha.v7.25326

20. Mishra VK, Retherford RD, Smith KR. Biomass cooking fuels and prevalence of tuberculosis in India. Int J Infect Dis. 1999;3: 119–129. Available: https://www.ncbi.nlm.nih.gov/pubmed/10460922

21. Kan X, Chiang C-Y, Enarson DA, Chen W, Yang J, Chen G. Indoor solid fuel use and tuberculosis in China: a matched case-control study. BMC Public Health. 2011;11: 498. doi:10.1186/1471-2458-11-498

22. Behera D, Aggarwal G. Domestic cooking fuel exposure and tuberculosis in Indian women. Indian J Chest Dis Allied Sci. 2010;52: 139–143. Available: https://www.ncbi.nlm.nih.gov/pubmed/20949731

23. Pakistan: Standard DHS, 2012-13 [Internet]. Available: https://dhsprogram.com/what-we-do/survey/survey-display-419.cfm

24. Lönnroth K, Williams BG, Cegielski P, Dye C. A consistent log-linear relationship between tuberculosis incidence and body mass index. Int J Epidemiol. 2010;39: 149–155. doi:10.1093/ije/dyp308

25. Cegielski JP, McMurray DN. The relationship between malnutrition and tuberculosis: evidence from studies in humans and experimental animals. Int J Tuberc Lung Dis. 2004;8: 286–298. Available: https://www.ncbi.nlm.nih.gov/pubmed/15139466

26. Qadeer E, Fatima R, Yaqoob A, Tahseen S, Haq MU, Ghafoor A, et al. Population Based National Tuberculosis Prevalence Survey among Adults (>15 Years) in Pakistan, 2010–2011. PLoS One. 2016;11: e0148293. doi:10.1371/journal.pone.0148293

27. TB Pakistan - Small Area Estimation [Internet]. Available: https://github.com/SCTX/tb_pakistan_sae

28. Initial data exploration of the prevalence surveys [Internet]. Available: https://github.com/SCTX/tb_pakistan_sae/blob/master/notebooks/1-cleaning/2-prevalence-survey-dataset/1.1-wt-prevalence-survey-data-initial-cleaning-and-consistency-check.ipynb

29. Histogram of Cases per Cluster in the Prevalence Survey Dataset [Internet]. Available: https://github.com/SCTX/tb_pakistan_sae/blob/master/notebooks/1-cleaning/2-prevalence-survey-dataset/1.4-wt-prevalence-survey-histogram-case-per-cluster.ipynb

30. run generalized linear models to estimate covariate effect on TB prevalence [Internet]. Available: https://github.com/SCTX/tb_pakistan_sae/blob/master/src/R/lib/make_prev_survey_data_glm_mods.R

31. SpatioTemporal Analysis and Mapping in Python [Internet]. Available:

https://github.com/InstituteforDiseaseModeling/STAMP

32. Clean and aggregate all the available Alhasan shapefiles [Internet]. Available: https://github.com/SCTX/tb_pakistan_sae/blob/master/notebooks/1-cleaning/1-pakistance-maps/1.3-wt-cleaning-and-aggregation-of-alhasan-shapefiles.ipynb

33. Covariate processing and testing files [Internet]. Available: https://github.com/SCTX/tb_pakistan_sae/tree/master/src/R/lib

34. Generic Code for running geostatistical models with INLA [Internet]. Available: https://github.com/rburstein-IDM/generic_mapper

35. World Health Organization. Tuberculosis prevalence surveys: a handbook. 2011.

36. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, Articles. 2015;67: 1–48. doi:10.18637/jss.v067.i01

37. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Series B Stat Methodol. 2009;71: 319–392. doi:10.1111/j.1467-9868.2008.00700.x

38. District-level TB prevalence estimation in Pakistan (clean scripts version) [Internet]. Available: https://github.com/SCTX/tb_pakistan_sae_clean

39. Burgert-Brucker CR, Dontamsetti T, Gething PW. The DHS Program's Modeled Surfaces Spatial Datasets. Stud Fam Plann. 2018;49: 87–92. doi:10.1111/sifp.12050

40. Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA [Internet]. Spatial and Spatio-temporal Epidemiology. 2013. pp. 33–49. doi:10.1016/j.sste.2012.12.001