SAE-LM modelling using TB prevalence survey data

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Fulvia Mecatti

University of Milano-Bicocca

TB MAC & WHO Annual Meeting 2019, Istanbul, 01-04 October

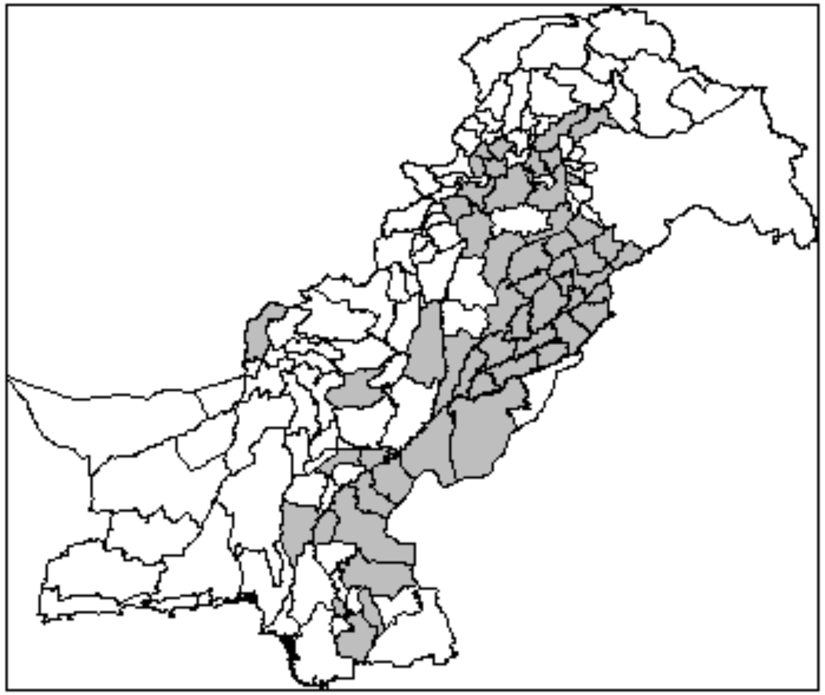
What *SAE-LM modelling* is

- *New* **methodology for decomposing a national estimate into sub-national estimates**
  PhD thesis 2016, paper on *ISR* 2018

- **Small Area Estimation integrated with Latent Markov modelling**

- **Showcase: Population based national TB prevalence survey among adults in Pakistan, Aug 2010 – Dec 2011**
  Pulmonary TB Bact+ $\geqslant 15$ years

Why a *SAE problem*

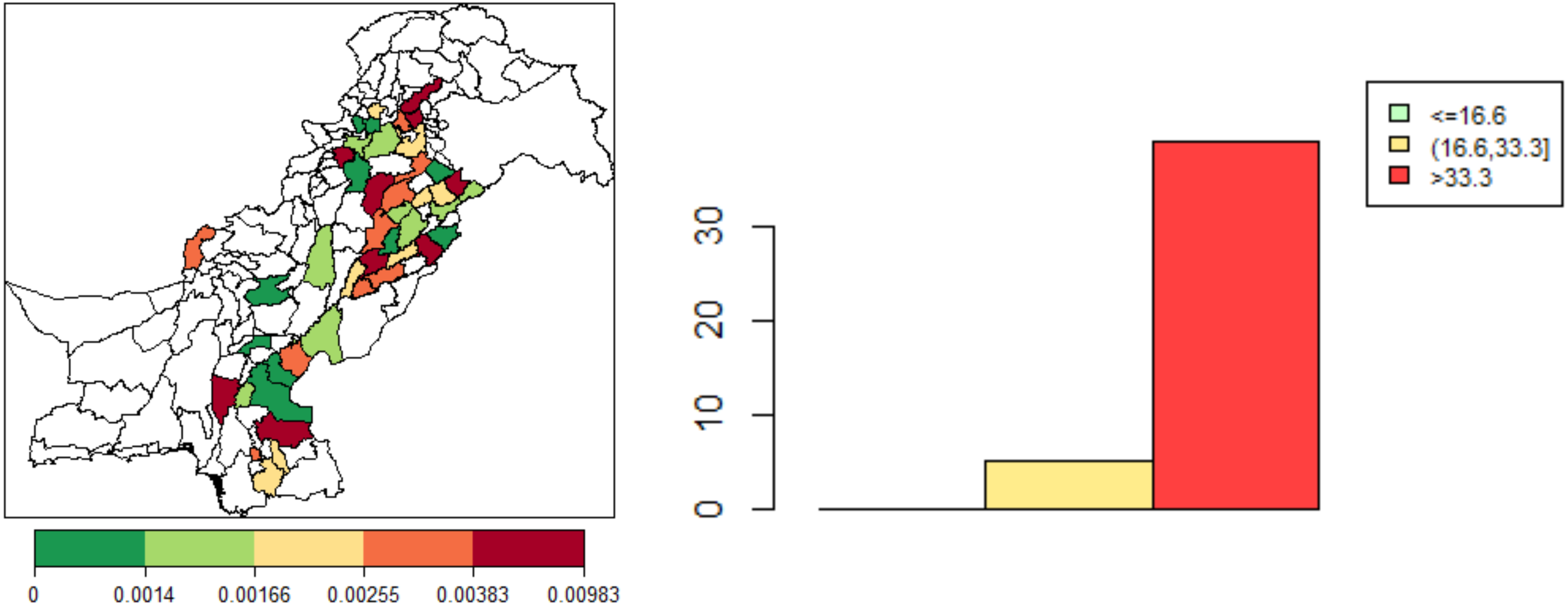


Sampled (gray) & non-sampled districts

- National survey designed to produce a national estimate with controlled precision **20% max error at 95% c.l.** upon a sample of **133,000** adult individuals from 95 clusters (thensils = sub-district areas) selected according to the RedBook guidelines.

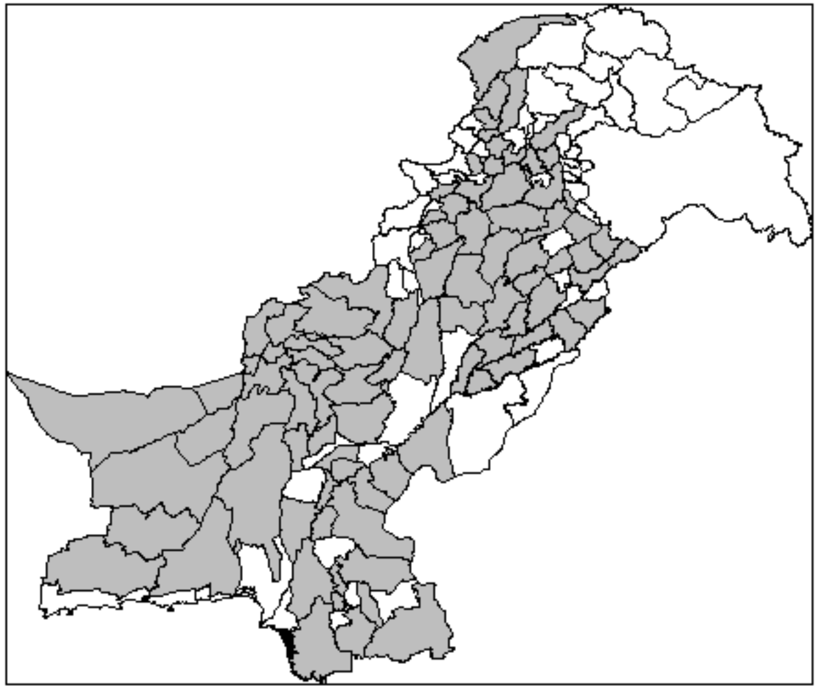
- **Random sample size** too small for accurate estimation of **district-wise** prevalence, limited to **68 districts** which either intersect or include at least one sampled cluster (thensil). Otherwise no sample data available

Why a *SAE problem*



Direct estimates for sampled districts and their CVs (%)

SAE-LM basic idea



districts with complete covariates

- To borrow straight from out-of-sample auxiliary data (model **covariates**)

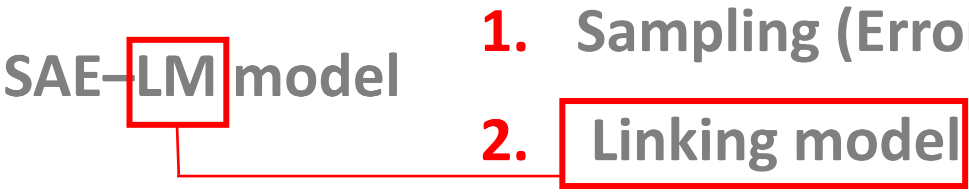
- Notification and Census auxiliary longitudinal data shared by Pakistan NTP and TB Hack team
  circa 30 covariates,   2011-2016 (2017)
  96 districts – 47 sampled, 49 zero sample size

SAE-LM Rational

- **Full exploit of both cross-sectional data (2010-11 national survey) and longitudinal auxiliary data (2011-2016 @district level)**

- **True district prevalence values as latent responses,** partially and indirectly measured at 6 successive time points, *i.e.* a underlying **Latent Process**

- **Distribution in space and evolution in time modelled** via a (discrete, 1° order) **Markov Chain** and a **Hierarchical Bayes** approach

- 

SAE–LM model

1. **Sampling (Error) model**
2. **Linking model**

SAE-LM at work (basics)

1. (SAE) **Sampling model** — Probability distribution of **Direct estimates** (input) conditioned on the true values of district prevalence

2. **Measurement model** — Probability distribution of the true values of district prevalence given the ***covariates*** (measurable part of the underlying latent process)

3. **Latent model** (Markov Chain) — will catch all the **residual heterogeneity** un-observed and un-explained

SAE-LM at work (basics)

- **Fitting** **Computational Algorithm: Data Augmentation Markov Chain Monte Carlo (Gibbs sampler)**
  60,000 MC runs, after 30,000 burn-in period for each combinations of (selected) subset of covariates and # of latent states

- **Selection** **Information Criterion based on maximum likelihood measure of goodness of fit**: **Chib's** (marginal likelihood) validated by **BIC** and **AIC** both in accordance

- **Validation** via several diagnostics tools

SAE-LM: Limits & Strengths

- **resource consuming**

  adequate statistical skills, computational power and time required

- **unfamiliar Bayesian approach to estimation**

  chosen over more familiar frequentist fixed parameter paradigm

- **Posterior Probability Distribution (MCMC Gibbs)**

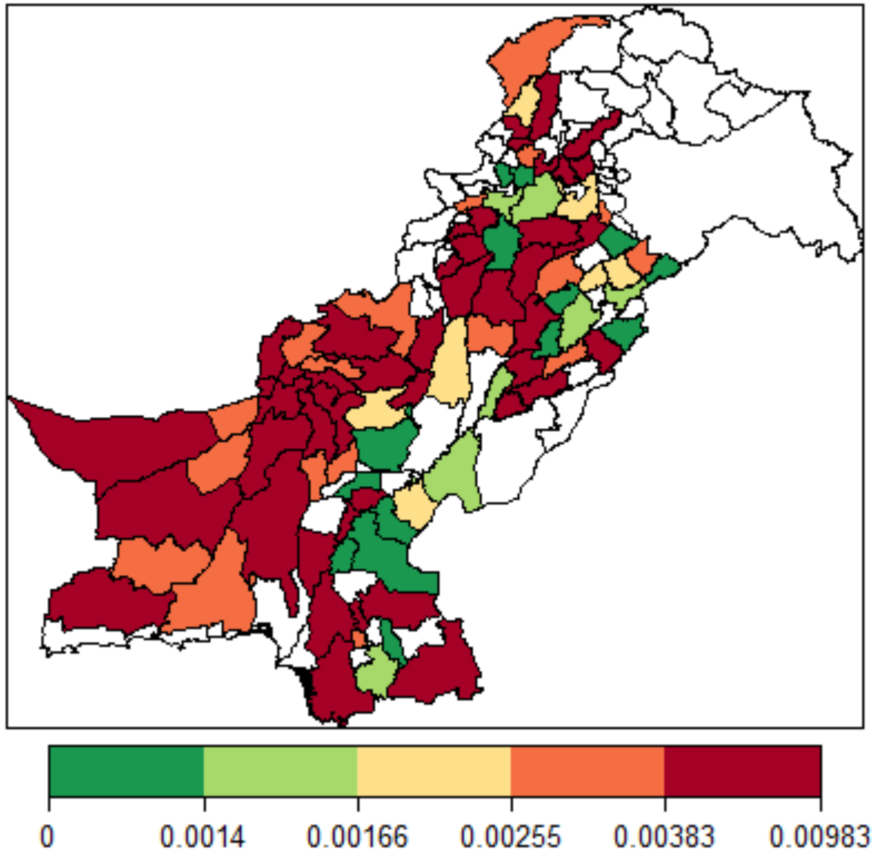
  Missing value Imputation (non-sampled districts no direct estimate in input)

  95% Credible Intervals

- **model Output**

  rich, easy to read, ready-to-use & includes a predictive engine

SAE-LM Results

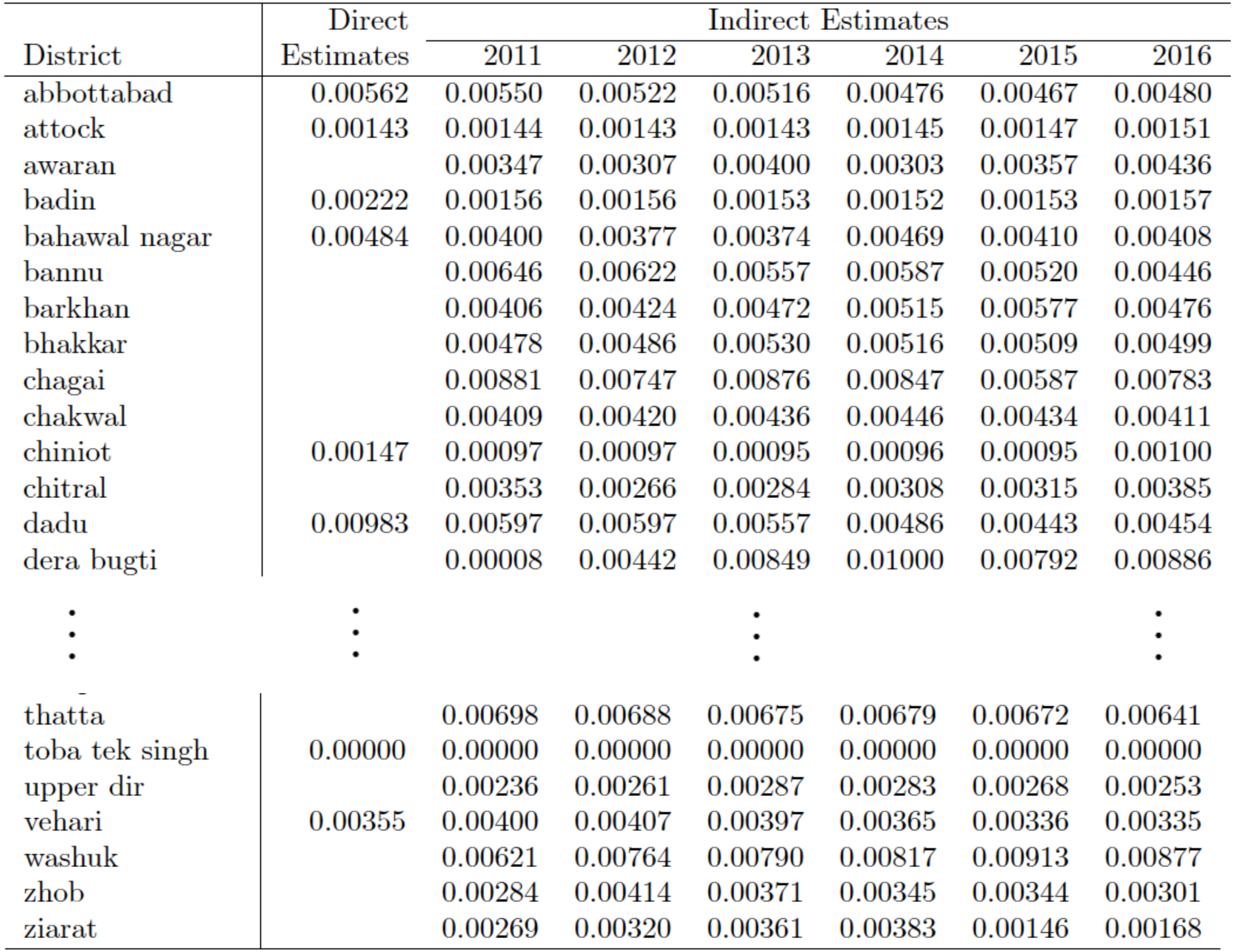


Sub-national prevalence estimates pulmonary TB Bact+ adults

- **District-wise TB prevalence (indirect, point) estimates (Distribution in space)**
- **Evolution in time of TB burden**

  Indirect estimates for each point in time

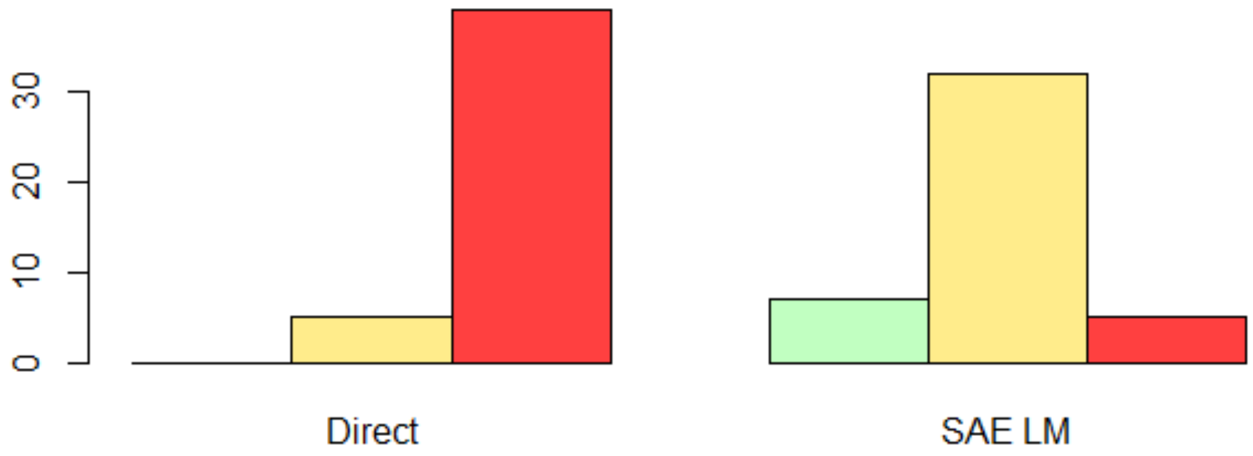
SAE-LM Results

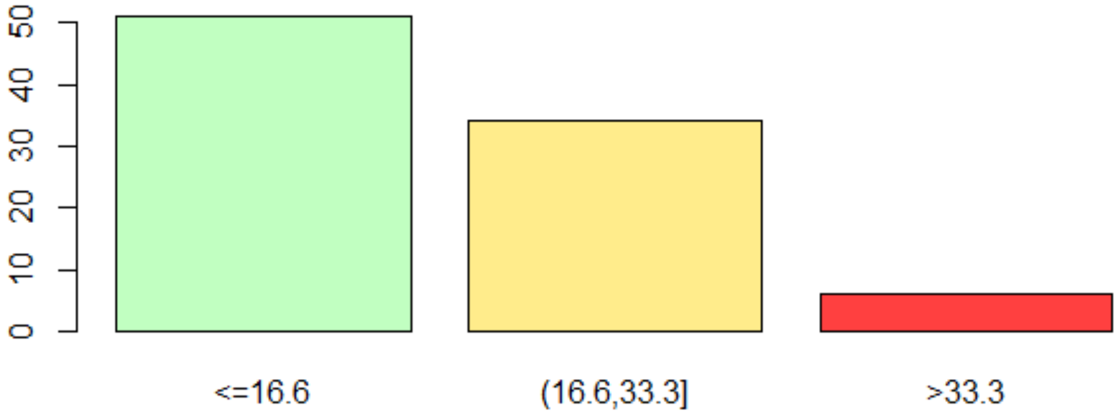


District	Direct Estimates	Indirect Estimates 2011	2012	2013	2014	2015	2016
abbottabad	0.00562	0.00550	0.00522	0.00516	0.00476	0.00467	0.00480
attock	0.00143	0.00144	0.00143	0.00143	0.00145	0.00147	0.00151
awaran		0.00347	0.00307	0.00400	0.00303	0.00357	0.00436
badin	0.00222	0.00156	0.00156	0.00153	0.00152	0.00153	0.00157
bahawal nagar	0.00484	0.00400	0.00377	0.00374	0.00469	0.00410	0.00408
bannu		0.00646	0.00622	0.00557	0.00587	0.00520	0.00446
barkhan		0.00406	0.00424	0.00472	0.00515	0.00577	0.00476
bhakkar		0.00478	0.00486	0.00530	0.00516	0.00509	0.00499
chagai		0.00881	0.00747	0.00876	0.00847	0.00587	0.00783
chakwal		0.00409	0.00420	0.00436	0.00446	0.00434	0.00411
chiniot	0.00147	0.00097	0.00097	0.00095	0.00096	0.00095	0.00100
chitral		0.00353	0.00266	0.00284	0.00308	0.00315	0.00385
dadu	0.00983	0.00597	0.00597	0.00557	0.00486	0.00443	0.00454
dera bugti		0.00008	0.00442	0.00849	0.01000	0.00792	0.00886
⋮	⋮			⋮			⋮
thatta		0.00698	0.00688	0.00675	0.00679	0.00672	0.00641
toba tek singh	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
upper dir		0.00236	0.00261	0.00287	0.00283	0.00268	0.00253
vehari	0.00355	0.00400	0.00407	0.00397	0.00365	0.00336	0.00335
washuk		0.00621	0.00764	0.00790	0.00817	0.00913	0.00877
zhob		0.00284	0.00414	0.00371	0.00345	0.00344	0.00301
ziarat		0.00269	0.00320	0.00361	0.00383	0.00146	0.00168

SAE-LM Results Validation

Comparing CVs of direct and indirect SAE-LM estimates for sampled districts





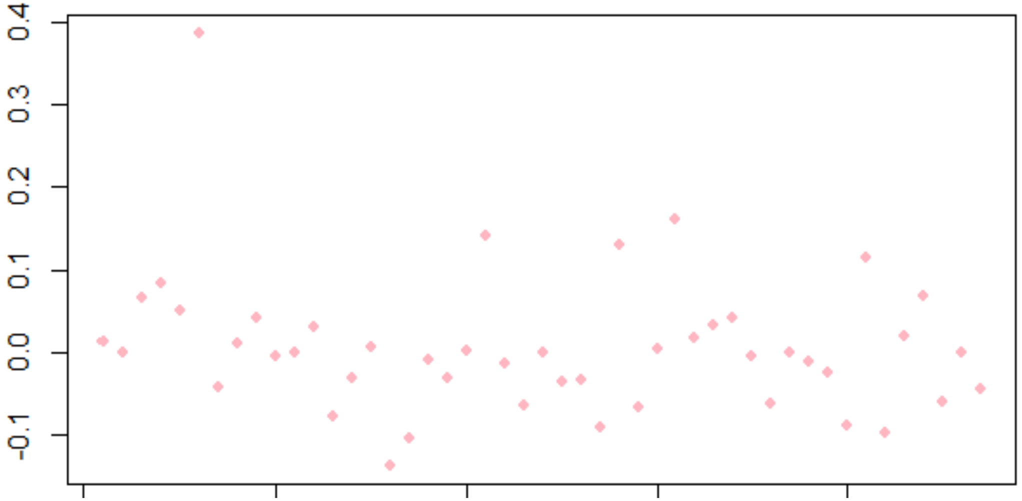
CVs of indirect SAE-LM district-wise estimates

SAE-LM Results Validation

The parts fit the whole

0.00361 average aggregation of the 96 district-wise indirect SAE-LM estimates

0.00364 national estimate 2010-11 Pakistan TB prevalence survey

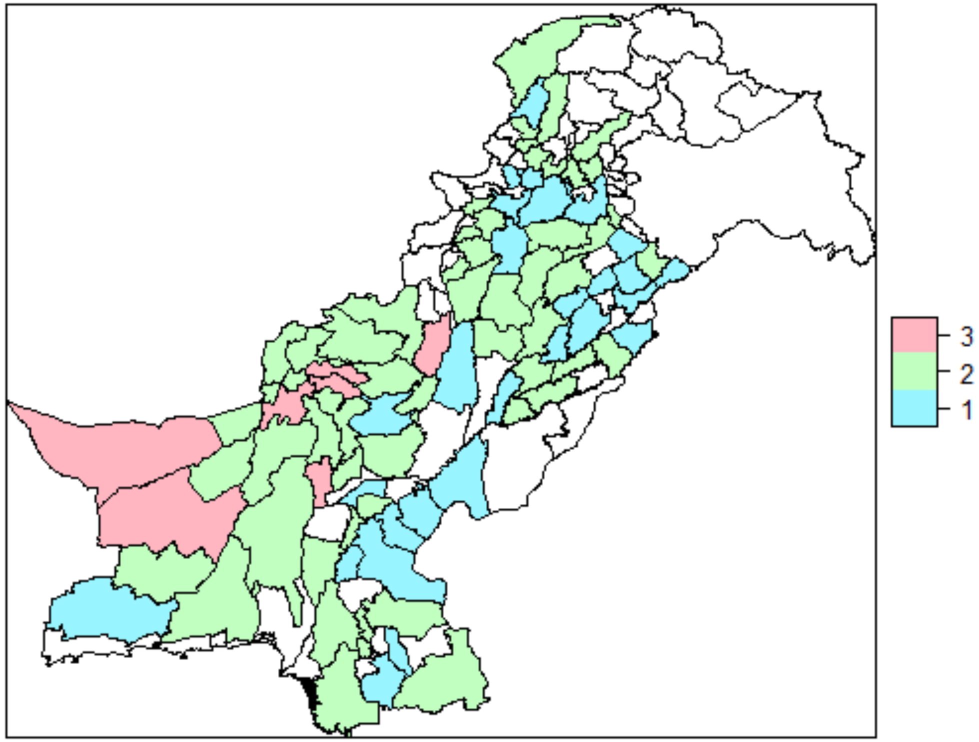


(Direct - Indirect SAE-LM)

vs

sampled districts sorted according to the variability (SE) direct estimates

SAE-LM Results

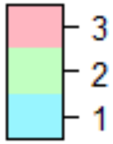


- **Classification** of Pakistan districts into three classes of increasing TB burden (the 3 latent states of the latent model)
- **Widely though not entirely reflecting the previous indirect estimates sub-national distribution:** some **relevant determinant(s)** of the district-distribution remain(s) **un-measured**, not included into the model covariates, so caught by the latent model as **residual heterogeneity**

SAE-LM Results

- **Initial** Probabilities

$$(0.322, 0.587, 0.091)$$

overall **average probability**, for a given district, to be classified into 

at time of national prevalence survey

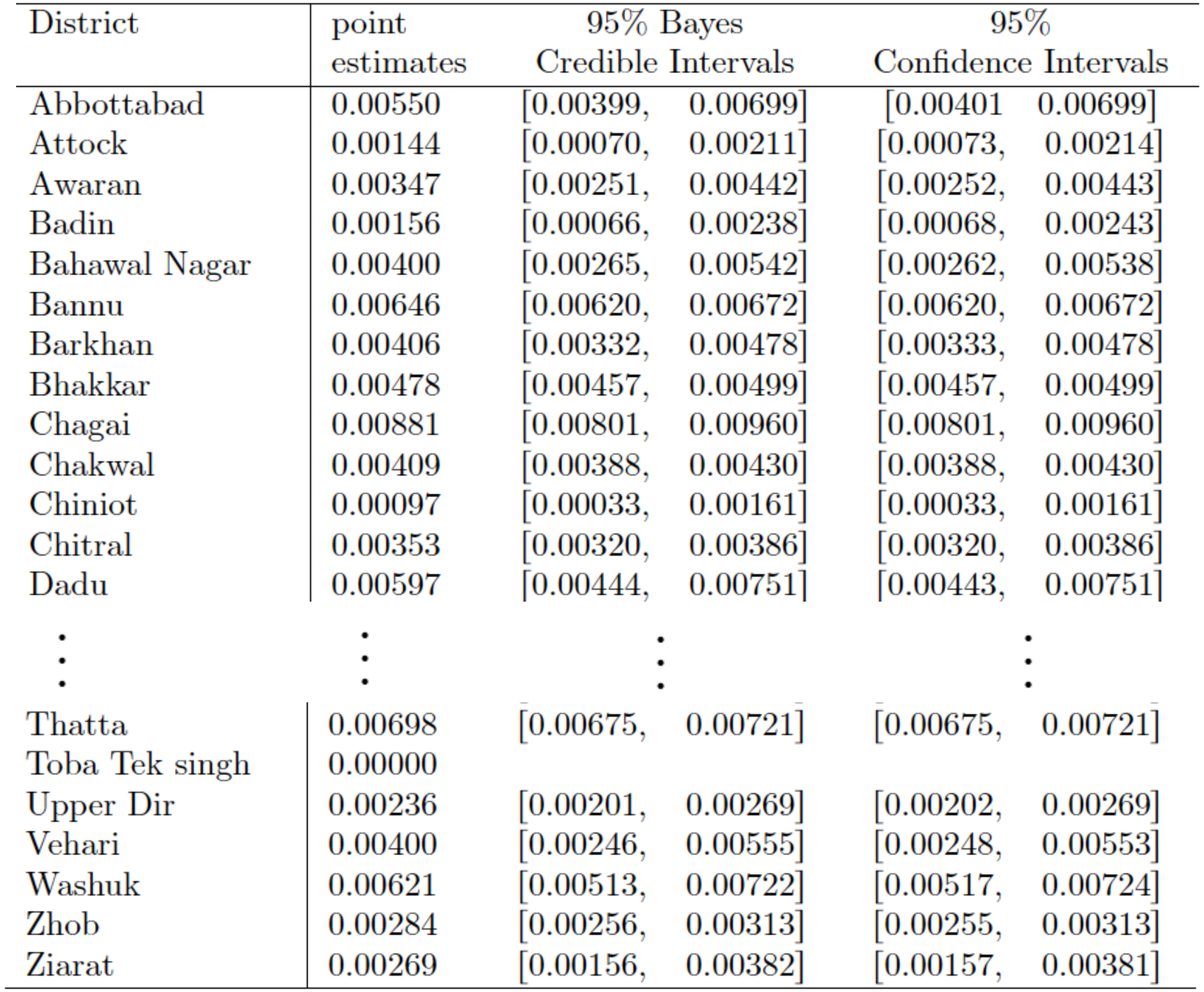
- **Transition** Probabilities

$$\begin{pmatrix} 0.975 & 0.012 & 0.014 \\ 0.005 & 0.984 & 0.010 \\ 0.021 & 0.033 & 0.946 \end{pmatrix}$$

for any district classified in a given class of TB burden (rows) will **predict** the chance for that district to **change class in the future**, moving backward (improving) or forward (worsening) across the matrix columns

Thank you

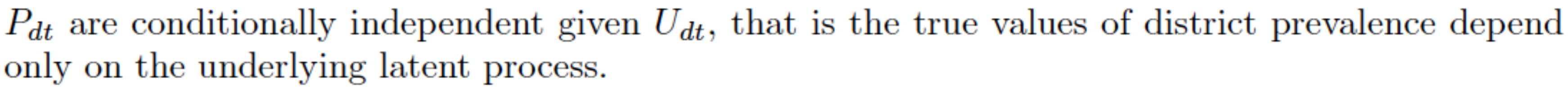
SAE-LM Uncertainty Intervals



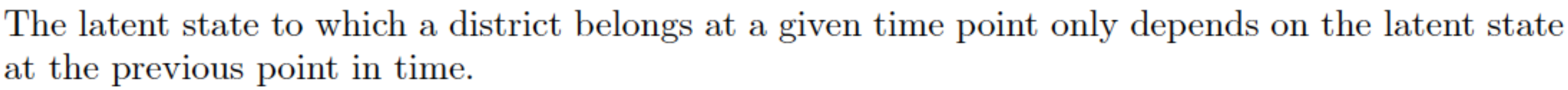
District	point	95% Bayes	95%
	estimates	Credible Intervals	Confidence Intervals
Abbottabad	0.00550	[0.00399, 0.00699]	[0.00401 0.00699]
Attock	0.00144	[0.00070, 0.00211]	[0.00073, 0.00214]
Awaran	0.00347	[0.00251, 0.00442]	[0.00252, 0.00443]
Badin	0.00156	[0.00066, 0.00238]	[0.00068, 0.00243]
Bahawal Nagar	0.00400	[0.00265, 0.00542]	[0.00262, 0.00538]
Bannu	0.00646	[0.00620, 0.00672]	[0.00620, 0.00672]
Barkhan	0.00406	[0.00332, 0.00478]	[0.00333, 0.00478]
Bhakkar	0.00478	[0.00457, 0.00499]	[0.00457, 0.00499]
Chagai	0.00881	[0.00801, 0.00960]	[0.00801, 0.00960]
Chakwal	0.00409	[0.00388, 0.00430]	[0.00388, 0.00430]
Chiniot	0.00097	[0.00033, 0.00161]	[0.00033, 0.00161]
Chitral	0.00353	[0.00320, 0.00386]	[0.00320, 0.00386]
Dadu	0.00597	[0.00444, 0.00751]	[0.00443, 0.00751]
⋮	⋮	⋮	⋮
Thatta	0.00698	[0.00675, 0.00721]	[0.00675, 0.00721]
Toba Tek singh	0.00000		
Upper Dir	0.00236	[0.00201, 0.00269]	[0.00202, 0.00269]
Vehari	0.00400	[0.00246, 0.00555]	[0.00248, 0.00553]
Washuk	0.00621	[0.00513, 0.00722]	[0.00517, 0.00724]
Zhob	0.00284	[0.00256, 0.00313]	[0.00255, 0.00313]
Ziarat	0.00269	[0.00156, 0.00382]	[0.00157, 0.00381]

SAE-LM model specification (basics)

- Main assumptions



P_{dt} are conditionally independent given U_{dt}, that is the true values of district prevalence depend only on the underlying latent process.



The latent state to which a district belongs at a given time point only depends on the latent state at the previous point in time.

SAE-LM model specification (basics)

- **Sampling (SAE) model**

direct district prevalence estimates given (conditioned on) true district prevalences

$$\hat{P}_d | P_d \sim N\left(P_d,\ \Sigma_d\right)$$

- **Measurement model**

true district prevalence given covariates, i.e. the measurable part of the latent process

$$P_{dt} | (U_{dt} = u) \sim N\left(\underline{x}_{dt}\underline{\beta}_u,\ \sigma_u^2\right)$$

- **Latent model**

probability distribution (discrete and dynamic in time) of the residual un-observed part of latent process, not explained by the upper hierarchy

$U_{dt} = u \sim$	(1st order) Markov chain with
k	latent states $u = 1, 2 \ldots k$
$[\pi_u]$	$k \times 1$ vector of initial probabilities
$[\pi_{u\|u'}]$	$k \times k$ matrix of transition probabilities, constant in time

SAE-LM model specification (basics)

- Small Area parameters $[P_{dt}]$ $D \times 6$ matrix of indirect area estimates
  primary output

- Measurement parameters $\left[\underline{\beta}_u, \ \sigma_u^2\right]$ $2k \times 1$ vector of regression coefficients and error variances

- Latent parameters $[\pi_u]$ k vector of initial probabilities,

  $[\pi_{u|u'}]$ $k \times k$ matrix of transition probabilities.