# TB-Fundi: an agent-based tuberculosis transmission model

Alvin X. Han and Colin A. Russell

**Summary**

TB-Fundi is an agent-based model that simulates the spread of *Mycobacterium tuberculosis* (*Mtb*) through a spatially- and age-structured population. Fundi means an expert or knowledgeable person, from the isiZulu and isiXhosa word for teacher, umFundisi. TB-Fundi follows a programmatic flow:

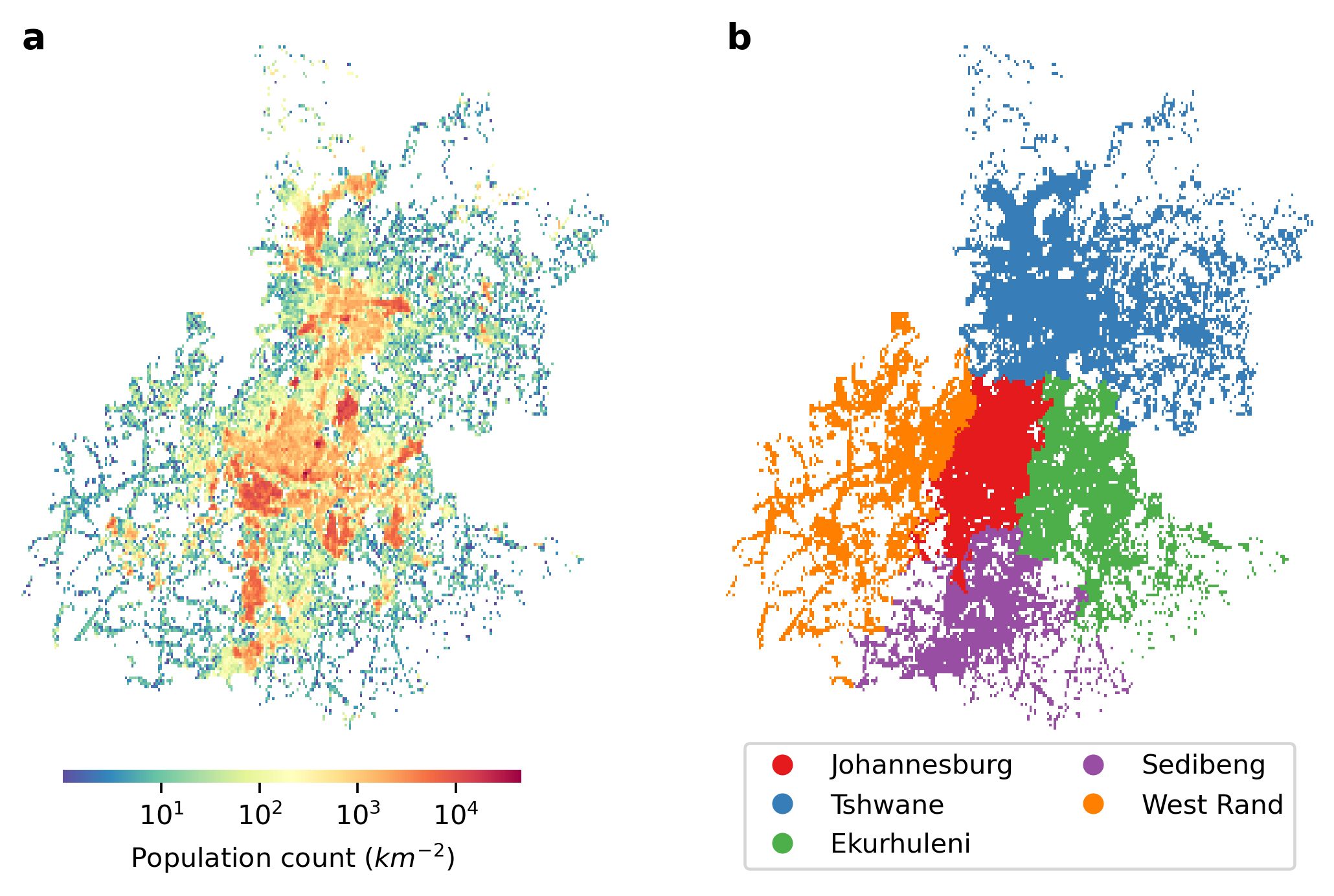
1. A population of simulated individuals is created based on demographic and location data. Contact networks between individuals are then set up based on their age and known population data (i.e. household composition data, school enrollment, employment rates, and employment firm size distribution) and human mobility properties.1
2. After setting up the population, the model iterates over a period of discrete time-steps (e.g. one-week), simulating the spread of the *Mtb* within the population created in step 1. During each time step, the simulation first updates the disease progression of each infected individual. Transmissions are then computed within households, schools, workplaces and through community contacts.

Source codes of TB-Fundi can be downloaded from <https://github.com/alvinxhan/tb-fundi>.

# Model details

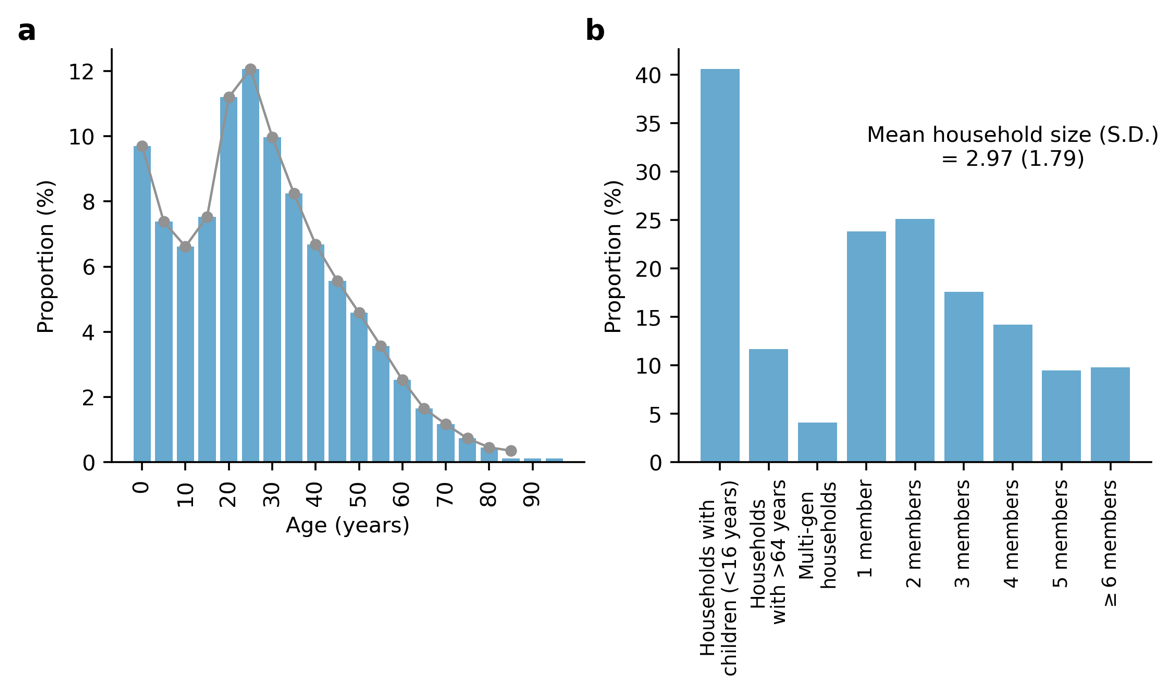
## Population

As a proof of concept, TB-Fundi was developed to simulate the transmission of Mtb in Gauteng province, South Africa. TB-Fundi first loads the spatially-distributed population count data (Fig. 1a) and subnational division of administrative units (Fig. 1b) for Gauteng from WorldPop.2 The population count data gives the estimated total number of people residing in every 1x1-km grid-cell while the subnational administrative units data (provided at 100x100-m grid-cell resolution) divides the grid-cells according to known administrative regions. Figure 1 shows the per-km grid-cell population density for the Gauteng in 2005 and the district (administrative region) to which each grid-cell belongs.



**Figure 1**: **Spatial distribution of population in Gauteng, South Africa in 2005 from WorldPop**.2 (**a**) Population count of each 1x1-km grid-cell. (**b**) District of each 1x1-km grid-cell.

## Households



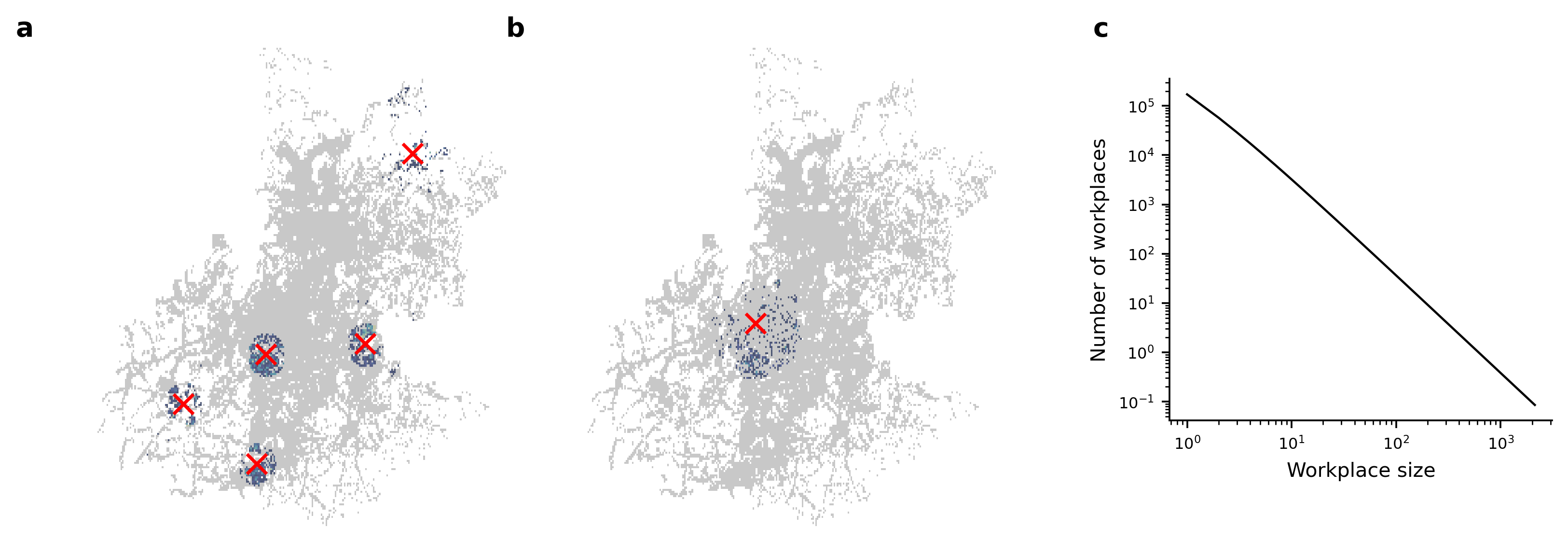
**Figure 2**: **Age demography and household composition of Gauteng, South Africa simulated by TB-Fundi**. (**a**) Bars show the population age distribution in 5-year bins simulated by TB-Fundi. Gray line shows the population age distribution based on 2011 population census data. (**b**) Simulated distribution of household composition for different types of households (with children <16 years of age; with senior adults >64 years of age; multigenerational-households; different household sizes). The reported mean household size in Gauteng in 2011 was 3 members per household.

TB-Fundi assigns the age of all individuals by multinomial sampling from age demography data (in 5-years bin) for each administrative region collected from population census (Fig. 2a).3 This is followed by sorting individuals into different households based on household size and composition data collated by the United Nations Population Division (Fig 2b).4 For each administrative region, TB-Fundi first calculates the expected number of households () by dividing the total population with the reported average household size () data. TB-Fundi then randomly sample individuals who would be the head of households based on reported age-stratified proportions. This is followed by calculating the expected size of households with multiple members () by:

where is the reported proportion of households with only one member from data. Among the sampled head of households, TB-Fundi randomly sample proportion of them to be the single-member in their respective households. TB-Fundi then randomly generate ) households, each assigned with the previously sampled head of household and a number of members drawn from a Poisson distribution with an expected number of members (which include the head of household as well). TB-Fundi then randomly sample a number of households with children <15 years of age based on the reported proportion in the data and randomly distributes all children of that age group into those households, ensuring that the Poisson-sampled number of members in those households is not exceeded. This is also done for household with adults >64 years of age. Finally, all remaining individuals are randomly distributed into multi-member households with remaining vacancies. Households (and as such, the individuals living in these households) are then randomly assigned locations on the grid-cells in each administrative region based on the probability that household of size will be located in grid-cell where number of individuals estimated to be residing in is:

where is a normalizing constant such that . Figure 2 shows the overall age demography and household composition of the population in Gauteng simulated by TB-FUNDI.

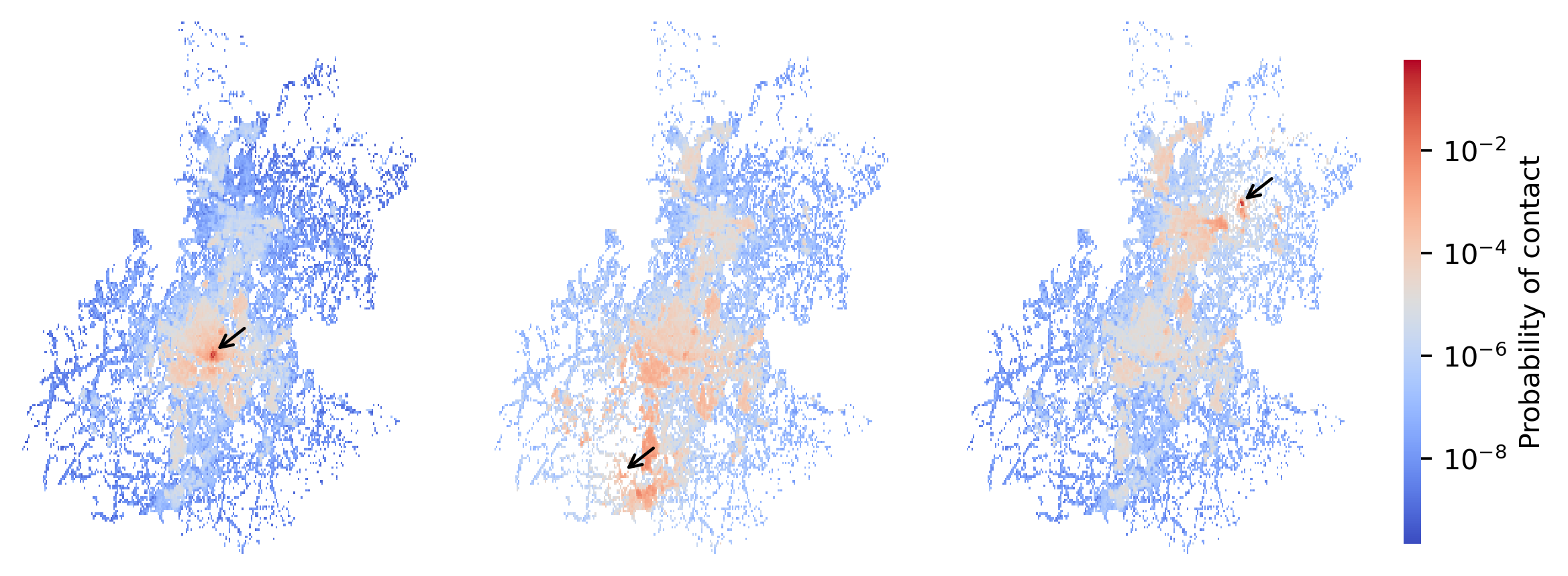
## Schools and Workplaces



**Figure 3**: **Simulated schools and workplaces**. (**a**) Example grid-cell locations of schools (marked as red crosses) in Gauteng. Blue dots surrounding a school (red cross) represent the grid-cell locations of students attending the school. (**b**) Example grid-cell location of a workplace (red cross) in Gauteng. Blue dots surrounding a workplace represent the grid-cell locations of employees employed at the workplace. (**c**) Assumed Zipf distribution with shape parameter for workplace employee sizes in Gauteng.

In TB-Fundi, schooling and employment are both explicitly simulated. Schools and workplaces are simulated in similar ways. For different schooling levels (i.e. primary and secondary) and workplaces in each administrative region, TB-Fundi first samples individuals of the stipulated age range (i.e. primary/secondary school age range or employment age range (e.g. 15-64 years)) living in the administrative region based on school enrollment and employment rate data. TB-Fundi requires data on the number of schools (i.e. and ; Fig. 3a) for each schooling level and the number of employment firms (; Fig. 3b) in each administrative region. and are often available from population census data. If data on are not available, TB-Fundi then assumes that the size of each firm follows the zipf distribution with the shape parameter (Fig. 3c).5 For Gauteng, TB-Fundi assumed that based on reported employment size of selected industries.6 TB-Fundi then randomly selects the grid-cell location for , and based on the population density of the grid-cells (i.e. ). Students and employees then “choose” the nearest school or workplace respectively based on the Haversine distance from their household location. As arbitrary upper limits, students will not “choose” a school further than 25-km from their homes while the distance between workplaces and homes will not exceed 50-km.

## Community Contacts



**Figure 4**: **Probability of random community contact between individuals residing in different grid-cell locations**. Each grid-cell is coloured by the probability of community contact () to individuals the index location marked by the arrow, as shown in the three example index locations.

To estimate the probability of random community contact between individuals residing in any two grid-cells within the simulated region, TB-Fundi assumes the universal visitation law of human mobility (Fig. 4).1 The probability that individuals living in location grid-cell visits and can thus be in contact with individuals living in grid-cell is:

where is the population density of location with radius , is the area of location , is the Haversine distance between location and and is the normalizing constant such that . is the frequency of visits individuals make to their home location and can be equal to 1 by assuming that all individuals return to their home location at least once a day. and are the respectively minimum and maximum number of frequencies to any location outside of their home location. is assumed to be (i.e. minimum visit for at least once per week) and is assumed to be 1 .

## Epidemic Initialization

TB-Fundi simulates the epidemic by first initializing infections in the region. This can be achieved by (1) randomly selecting susceptible individuals to be infected at timestep , (2) initializing a distribution of individuals in various disease states (see *Disease Progression*, *Model Calibration* and *Simulation* sections below), or (3) simulating a constant rate of external introductions of *Mtb* into the region. The probability that a susceptible individual would be infected by external introduction at timestep () is given by:

where is the external attack rate (i.e. number of expected infection due to external introduction per person per timestep) and is the fraction of susceptible individuals at timestep .

## Disease Progression



**Figure 5**: **Disease progression of *Mycobacterium tuberculosis* infection**.

TB-Fundi assumes that an individual infected with *Mtb* can be in different disease states (Fig. 5),7 including (1) *infection*, which is assumed to be those with evidence of *Mtb* by immunologic testing only but without clinical signs or symptoms of tuberculosis (TB); (2) *minimal* disease, which include individuals with pathology prior to onset of bacteriological evidence of TB regardless of symptoms; (3) *subclinical* disease refers to those with bacteriological evidence of TB but do not report TB symptoms; (4) *clinical* disease includes individuals with bacteriological evidence of TB and present symptoms of TB. Individuals with *subclinical* or *clinical* disease are infectious.

Some individuals in either the *infection* or *minimal* states can effectively control or eliminate their *Mtb* infection and *recover* without progressing into *subclinical*/*clinical* TB. *Recovered* individuals can also be re-susceptibilized and be reinfected at later timepoints. Several rate parameters are required by TB-Fundi to characterize the rate of infected individuals transitioning between different disease states (see Table 1 for example rates).

At the start of each timestep, TB-Fundi will update the disease state of infected individuals based on the transition probabilities as inferred from these rates. Suppose an infected individual is currently in state , they will transition to state with the probability given as:

where is the rate parameter going from state to state , is the duration that the individual has been in state .

**Table 1**: **Rate parameters between different disease states**.

|  |  |  |
| --- | --- | --- |
| **Parameter ()** | **Value** | **Source** |
| Recovered to Susceptible, | 0.87 per year | 8,9 |
| Infection to Recovered, | 1.83 per year | 7 |
| Infection to Minimal, | 0.10 per year | 7 |
| Infection to Subclinical, | 0.04 per year | 7 |
| Minimal to Recovered, | 0.18 per year | 7 |
| Minimal to Subclinical, | 0.24 per year | 7 |
| Subclinical to Minimal, | 1.58 per year | 7 |
| Subclinical to Clinical, | 0.72 per year | 7 |
| Clinical to Subclinical, | 0.57 per year | 7 |
| Clinical to Death, | 0.33 per year | 7 |

## Transmission

Upon contact with infectious individuals (i.e. those that are either in the subclinical or clinical disease state) at each time step , a susceptible individual of age group living in household , which is located in grid-cell , who is either studying at school or working at workplace , would be infected with the probability given as:

where is the per-unit-time risk of infection in the respective setting.

for home, school and workplaces are defined as:

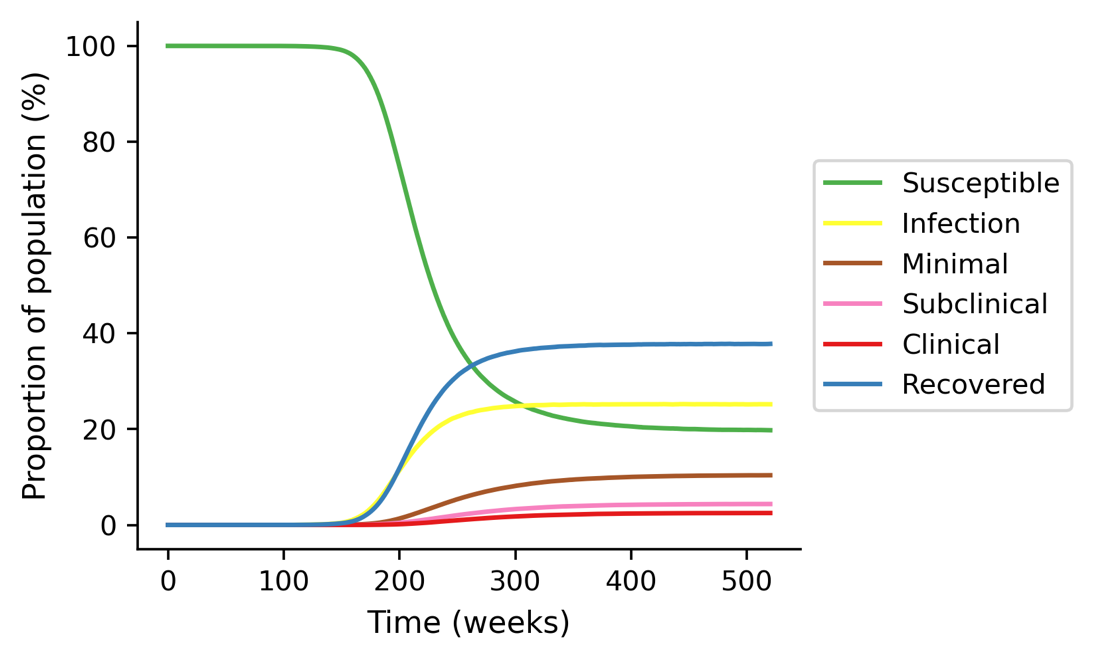
where is the transmission probability per contact, is the setting-specific mean number of contacts between individuals of age group and per unit time, is a Boolean variable indicating if member of age group in household or school or workplace is infected and is the number of individuals linked to those places (i.e. either living in household or attending school or working at workplace ). has been estimated by Prem et al. for various settings (i.e. household, school, workplace and community) in different countries.10

, on the other hand, is defined as:

where is the probability of contact between a susceptible individual who is living in location grid-cell and an infector who is living in location grid-cell .

# Model Usage

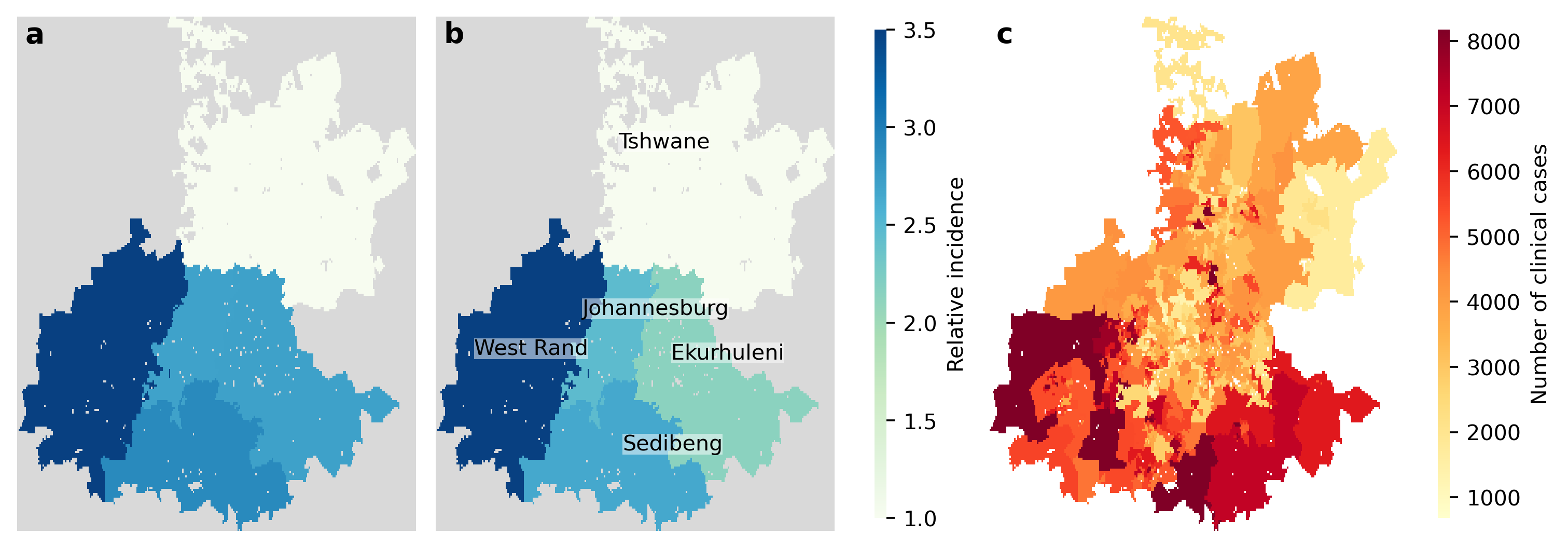
## Model Calibration



**Figure 6**: **Calibration TB-Fundi simulation over 10 years (520 weeks) in Gauteng assuming no Mtb-associated deaths to reach a “steady-state” distribution of disease states**.

Given the lack of data on prevalence stratified across the different disease states, a calibration simulation for a long period of time (e.g. 10 years or more), without the death state (i.e. population size remains constant) is recommended to estimate the initial “steady-state” distribution of states and time distribution individuals have stayed in those states (Fig. 6). The “steady-state” distribution provides an estimation of the current distribution of disease state across the population in geographical regions with a long history of *Mtb* circulation.

## Simulation Results



**Figure 7**: **Comparison between simulated clinical tuberculosis incidence and reported bacteriologically confirmed incidence rates in 2005**. (**a**) Simulated clinical incidence rates (per 100,000 people) relative to those in the City of Tshwane. (**b**) Reported relative bacteriologically confirmed incidence rates in 2005 relative to those in the City of Tshwane**.**11(**c**) Number of simulated clinical cases in each sub-district administrative region.

We simulated the spread of *Mtb* in Gauteng for one year, including deaths, initializing the epidemic from the “steady-state” disease state distribution (Fig. 6) as tabulated in Table 2. Individuals in the clinical state were randomly distributed across Gauteng, ensuring that relative incidence between districts was similar to reported bacteriologically confirmed incidence rates in 2004 (i.e. 501 per 100,000 people in City of Johannesburg, 149 per 100,000 people in City of Tshwane, 522 per 100,000 people in Ekurhuleni, 603 per 100,000 people in Sedibeng and 734 per 100,000 people in West Rand).11 As a preliminary assessment of TB-Fundi, we compared the relative simulated clinical incidence rates between different districts (Fig. 7a) to relative bacteriologically confirmed incidence rates reported in 2005 (Fig. 7b; 708 per 100,000 people in City of Johannesburg, 288 per 100,000 people in City of Tshwane, 615 per 100,000 people in Ekurhuleni, 763 per 100,000 people in Sedibeng and 1023 per 100,000 people in West Rand).11

**Table 2**: **Initial state distribution of population and time period in state**

|  |  |  |
| --- | --- | --- |
| **State** | **Proportion of population** | **Time period in state** |
| Infection | 25.2% | Mean = 3.6 weeks;  S.D. = 3.1 weeks |
| Minimal | 10.4% | Mean = 8.4 weeks;  S.D. = 6.7 weeks |
| Subclinical | 4.4% | Mean = 3.3 weeks;  S.D. = 2.9 weeks |
| Clinical | 2.5% | Mean = 7.1 weeks;  S.D. = 5.8 weeks |
| Recovered | 37.8% | Mean = 5.7 weeks;  S.D. = 4.7 weeks |

Similar to the reported data, the City of Tshwane had the lowest incidence with an estimated 2,835 cases per 100,000 people in the simulation results. As there was no data on testing rates that gave rise to the reported incidence in 2005 and transition rate parameters between compartmental states (Table 1) had been differently estimated in other studies,12,13 it is difficult to directly compare the estimated incidence from simulations to those reported in 2005. To semi-quantitatively demonstrate that TB-Fundi can be used to reflect the distribution of *Mtb* circulation in Gauteng, we computed the incidence rate in each district in Gauteng relative to those found in the City of Tshwane (Fig. 7a-b). The simulation results are largely in line with the reported relative incidence rates: the largest incidence rates were found in West Rand (i.e. simulation, 3.3 times relative to estimated incidence in the City of Tshwane; data, 3.6 times relative to confirmed incidence in the City of Tshwane), followed by Sedibeng (i.e. simulation, 3.0 times relative to estimated incidence in the City of Tshwane; data, 2.6 times relative to confirmed incidence in the City of Tshwane).

Taken together, the results presented here suggest that TB-Fundi is capable reproducing high-level spatial patterns of observed *Mtb* infections. With more robust and detailed data, TB-Fundi could be appreciably better parameterized and used to investigate more detailed patterns *Mtb* incidence and transmission.

# References

1. Schläpfer, M. *et al.* The universal visitation law of human mobility. *Nature* **593**, (2021).

2. WorldPop (www.worldpop.org - School of Geography and Environmental Science, U. of S. D. of G. and G. U. of L. D. de G. U. de N. & Center for International Earth Science Information Network (CIESIN), C. U. Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076). https://dx.doi.org/10.5258/SOTON/WP00670 (2018).

3. Statistics South Africa. Census 2011. https://www.statssa.gov.za/?page\_id=3836.

4. United Nations Population Division. Household Size and Composition. https://www.un.org/development/desa/pd/data/household-size-and-composition (2022).

5. Axtell, R. L. Zipf Distribution of U.S. Firm Sizes. *Science (1979)* **293**, 1818–1820 (2001).

6. Statistics South Africa. Annual Financial Statistics 2011. https://www.statssa.gov.za/publications/P0021/P00212011.pdf (2011).

7. Horton, K. C., Richards, A. S., Emery, J. C., Esmail, H. & Houben, R. M. G. J. Reevaluating progression and pathways following Mycobacterium tuberculosis infection within the spectrum of tuberculosis. *Proc Natl Acad Sci U S A* **120**, e2221186120 (2023).

8. Uys, P. *et al.* The risk of tuberculosis reinfection soon after cure of a first disease episode is extremely high in a hyperendemic community. *PLoS One* **10**, (2015).

9. Andrews, J. R. *et al.* Risk of progression to active tuberculosis following reinfection with Mycobacterium tuberculosis. *Clinical Infectious Diseases* **54**, (2012).

10. Prem, K. *et al.* Projecting contact matrices in 177 geographical regions: An update and comparison with empirical data for the COVID-19 era. *PLoS Comput Biol* **17**, e1009098 (2021).

11. National Institute for Communicable Diseases, S. A. *Microbiologically Confirmed Tuberculosis 2004-15 South Africa*. https://www.nicd.ac.za/wp-content/uploads/2019/11/National-TB-Surveillace-Report\_2004\_2015\_NICD.pdf (2017).

12. Dodd, P. J. *et al.* Transmission modeling to infer tuberculosis incidence prevalence and mortality in settings with generalized HIV epidemics. *Nat Commun* **14**, (2023).

13. Ku, C. C. *et al.* Durations of asymptomatic, symptomatic, and care-seeking phases of tuberculosis disease with a Bayesian analysis of prevalence survey and notification data. *BMC Med* **19**, (2021).