S1 Text. Simulation Study.

We carried out a simulation study to assess the predictive ability of geostatistical models that combine both point-referenced and areal survey data, by setting the locations of the areas (e.g., counties) at their centroids (i.e., model 1) compared to models that include only point-referenced data (i.e., model 2) for analysis.

We avoided high computation burden, by considering in the simulation only two neighboring provinces, Guangdong and Guangxi. 30 datasets were generated following the structure of the survey data in the region. We randomly sampled 12 points within each survey county which we treated them as the “unknown” survey locations in the county. A Gaussian process was generated over the whole set of locations (both observed and “unknown”), that is $w=(w\_{1},…,w\_{s})^{T}$, $w\~MVN(0,Σ)$, and $Σ\_{ij}=σ^{2}exp⁡(-ρd\_{ij})$, where $d\_{ij}$ is the Euclidean distance between locations $i$ and $j$. We set $σ^{2}=2$ and $ρ=5$. We considered a single covariate $X$ corresponding to the elevation and simulated survey data (i.e., number of infected individuals at location $i$) from a binomial distribution $y\_{i}\~Bin\left(p\_{i},n\_{i}\right)$ with binomial proportion $p\_{i}=logit^{-1}(β\_{0}+β\_{1}X\_{i}+w\_{i})$ by setting $β\_{0}=-2$ and $β\_{1}=-2$ and by assuming equal number of screened individuals $n\_{i}=500$ across locations. The simulated areal data was calculated as $Y\_{k}=\sum\_{i\in A\_{k}}^{}y\_{i}$ and $N\_{k}=\sum\_{i\in A\_{k}}^{}n\_{i}$, where $Y\_{k}$, $N\_{k}$ and $A\_{k}$ represent the number of infected, number of screened individuals and all the sampled points in county $k$. Simulations were carried out in R version 3.3.2.

The model fitting procedure is the same to the one presented in the statistical analysis section in the main text. We randomly selected 20% of point-referenced and of sampled within areas points for model validation. The remaining ones were for model fit. The mean absolute error (MAE) and predictive logarthmic score [1] were used for model validation. Lower MAE and higher logarithmic score suggest a model with better predictive ability.

The results from the simulation study are shown in Supporting Information S2 Fig. For better visualization, we connected datasets of the same model with lines in S2 Fig A and C. Model 1 had lower MAE and higher log score than that of model 2, suggesting a better predictive ability.

Reference

 1. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007; 102: 359-378.