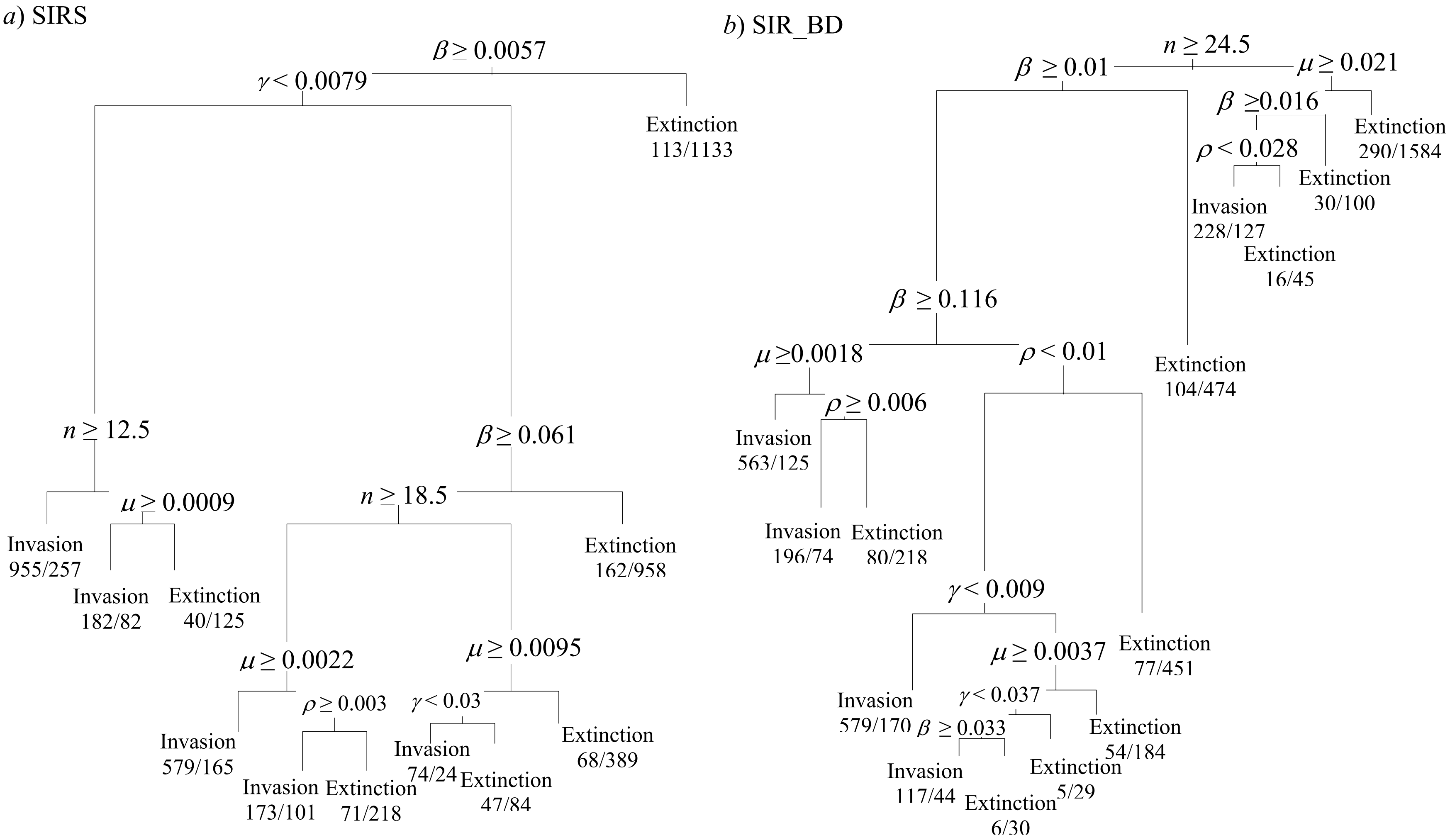
Electronic Supplementary Material
Additional details on classification trees and SIR_BD model results

Here we present alternative classification trees to those shown in the main text as well as the results of the SIR_BD model. During the classification tree analysis we ran three different analyses for each of the SIRS and SIR_BD models. Our first analysis of each model used only the raw parameter values as predictor variables (β, γ, μ, ρ, δ, and n). While the predictive power of this approach was useful, we found the trees less helpful as a heuristic tool because the absolute value of many parameters is meaningless without reference to other parameters in that system (Fig. E1). As a result, the trees are larger than those in the main text in an attempt to account for the importance of these interactions only the raw parameters rather than the ratios of different parameters.

In the second analysis, we calculated aggregate parameters that were more likely to illustrate the critical dependence among parameters. The aggregate parameters for the SIRS model were: β/γ, $\mu n/\gamma$, ρn, $\rho n/\gamma$. The ratio of β to γ represents the traditional R_0 for this model and $\mu n/\gamma$ is a rough approximation of the expected number of infectious migrants per group when R_0 is large. We tested three approaches to including the loss of immunity. First, the probability of losing immunity multiplied by the total group size should relate to the total rate at which new susceptibles enter the population. Secondly, we hypothesized that ρn may need to scaled by the infectious period ($1/\gamma$), similar to the movement rate. Finally, we allowed ρ to enter the model individually. In almost all analyses, ρn and ρ were not good predictor variables and were dropped from the classification trees. The classification trees using the aggregate parameters were the most informative. For clarity, in the main text (Fig. 3a) we show the best classification trees (as selected by the 1-SE rule; Breiman 1984) that used each predictor variable no more than once. The trees that resulted in the least amount of cross-validation error allowed R_0 to appear twice (Fig. E2), but were only marginally better than those described in the main text (Table 1). We also simulated the disease model using non-spatial movement rules where individuals could move to any other group in the array. The non-spatial movement rule resulted in classification trees very similar to those of the nearest-neighbour model, but the threshold level of movement $\mu n/\gamma$ was slightly reduced (Fig. 3a and Fig. E3).

In the third classification analysis, we used ν and ν_* as predictor variables along with the aggregate parameters described above (see Methods for a description of how ν and ν_* are estimated empirically from the simulations). The quantities ν and ν_* can only be calculated once each epidemic has run its course, in contrast to the other parameters which can be calculated a priori. We included ν and ν_* in this analysis as a test of 'gold-standard' predictor variables, and to formalize our findings from Fig. 1 within the classification tree framework. Although this analysis highlights ν_* as the strongest predictor of disease invasion in a metapopulation (Fig. E4, Table 1), it is only once an epidemic is well underway that this value can be calculated in principle, and the logistical barriers to collecting the relevant contact tracing data are formidable. Thus ν and ν_* include information on the vitally important stochasticity of the initial stages of invasion (in essence, they 'know how the dice fell'), and 'predictions' based on ν and ν_* result in far fewer false-positives.



a) SIRS

b) SIR_BD

Figure E1. Classification trees for the SIRS (a) and SIR_BD models (b) using only the raw model parameters to explain disease invasion in a metapopulation. Threshold criteria are labeled above each node of the tree, and instances that satisfy the criteria are split off to the left. Labels underneath the terminal leaves indicate the number of simulations (out of 6000) resulting in invasions and extinctions, respectively, and in text the majority outcome for that set of classification rules.

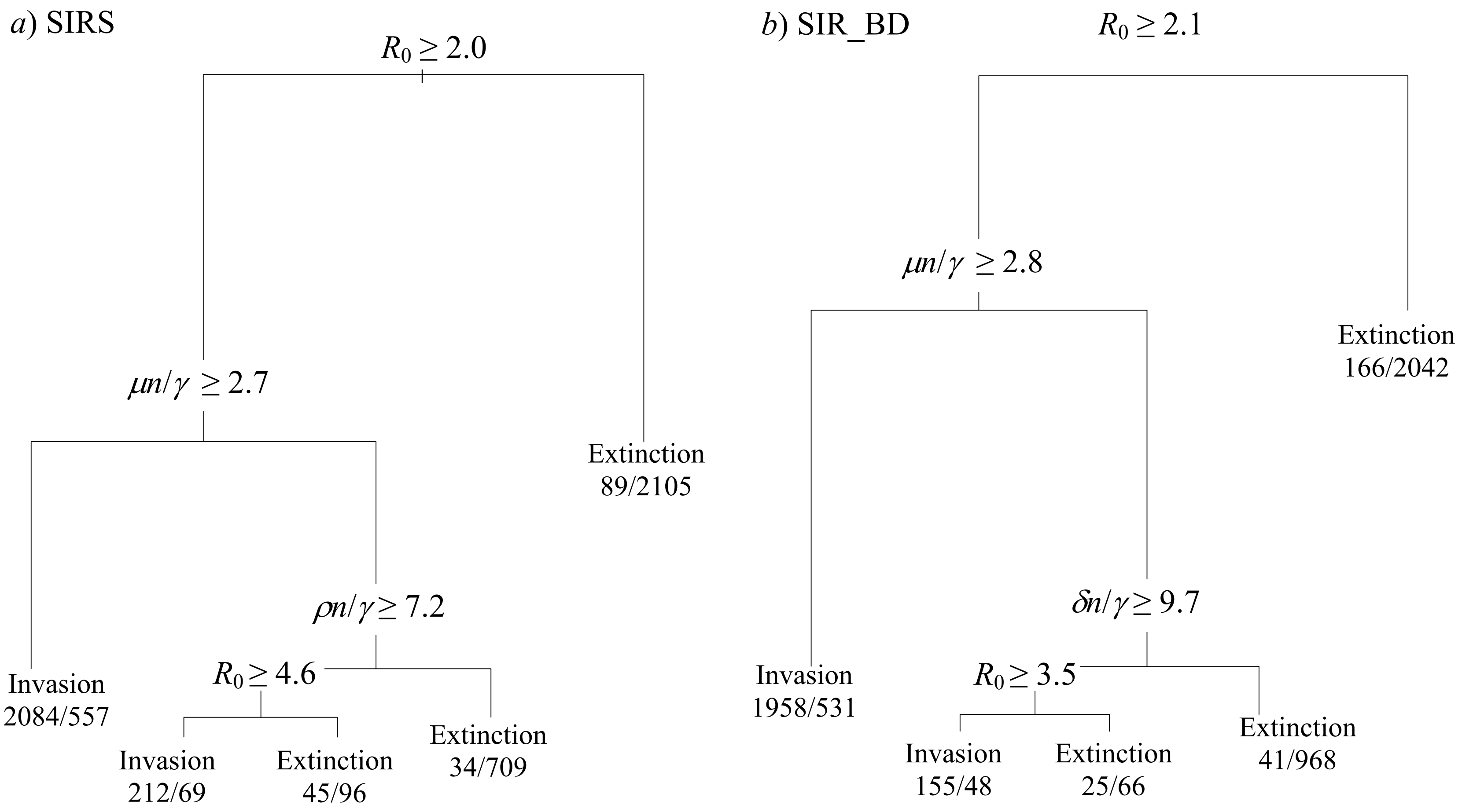


Figure E2. 'Best'classification trees as determined by the 1-SE rule for the SIRS (a) and SIR_BD models (b) using aggregated model parameters to explain disease invasion in a metapopulation.

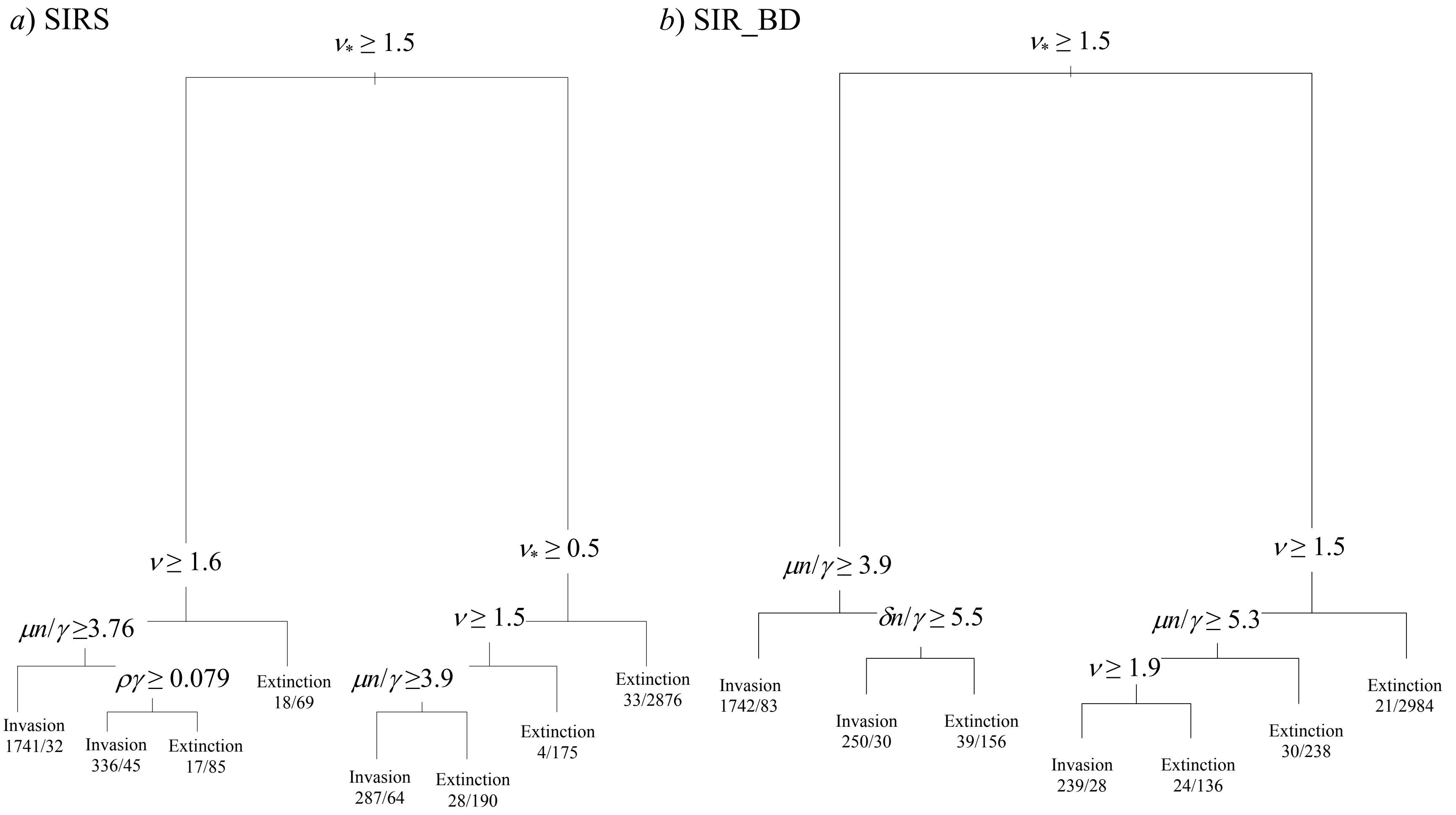


Figure E3. 'Best'classification trees as determined by the 1-SE rule for the SIRS (a) and SIR_BD models (b) using aggregated model parameters as well as the ν and ν_* metrics to explain disease invasion in a metapopulation.

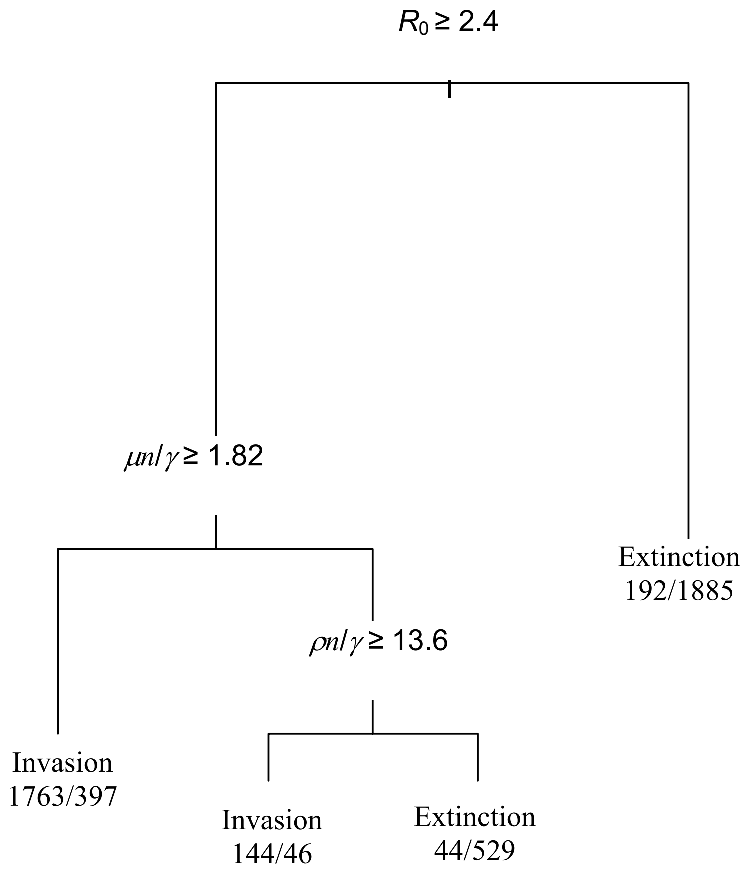


Figure E4. Classification trees for the SIRS non-spatial model. Threshold criteria are labeled above each node of the tree, and instances that satisfy the criteria are split off to the left. Labels underneath the terminal leaves indicate the number of simulations (out of 5000) resulting in invasions and extinctions, respectively, and in text the majority outcome for that set of classification rules.

Table E1. The proportion of SIR_BD model simulations where the disease invades the metapopulation and whether that invasion was predicted by theoretical thresholds or the classification tree analyses.

Rules for invasion	Correctly predicted invasions	Correctly predicted extinctions	False-positive[1]	False-negative[2]	Total misclassified	Cross-validated misclassification[3]	SD[3]
$R_0 > 1$	0.391	0.172	0.438	0	0.438	--	--
$\mu n/\gamma > 1$	0.377	0.182	0.427	0.014	0.441	--	--
$R_0 > 1$ and $\mu n/\gamma > 1$	0.377	0.336	0.273	0.014	0.287	--	--
best classification tree[4]	0.352	0.513	0.097	0.039	0.135	0.138	0.0045
reduced classification tree[5]	0.356	0.502	0.108	0.035	0.142	0.143	0.0045
raw parameter tree[6]	0.281	0.519	0.090	0.110	0.200	0.224	0.0054
$\nu > 1$, *post-hoc*[7]	0.340	0.426	0.184	0.051	0.234	--	--
$\nu_* > 1$, *post-hoc*[7]	0.339	0.564	0.045	0.052	0.097	--	--

[1] Rules predicted invasions when the disease actually went extinct.

[2] Rules predicted extinctions when the disease actually invaded.

[3] Average and standard deviation of error rates on test data not used in the construction of the classification tree using 10-fold cross-validation.

[4] Using aggregate parameters not including ν and ν_*. See Figure E2.

[5] Using the aggregate parameters not including ν and ν_*. See Figure 3.

[6] Using raw parameters not including ν and ν_*. See Figure E1.

[7] ν and ν_* are considered post-hoc because they can only be estimated after the epidemic has begun and thus have an advantage over other metrics included in the table