Title: An approach for modeling cross-immunity of two strains, with application to variants of Bartonella in terms of genetic similarity

Authors: Kwang Woo Ahn[a,*], Michael Kosoy[b], and Kung-Sik Chan[c]

Affiliation: [a]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI USA

[b]Centers for Disease Control and Prevention, Fort Collins, CO USA

[c]Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA USA

[*]Corresponding author at Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI 53226 USA. Tel.: +1 414 955 7387; fax: +1 955 6513

E-mail addresses: kwooahn@mcw.edu (K. W. Ahn), mck3@cdc.gov (M. Kosoy), kung-sik-chan@uiowa.edu (K. S. Chan)

Abstract

We developed a two-strain susceptible-infected-recovered (SIR) model that provides a framework for inferring the cross-immunity between two strains of a bacterial species in the host population with discretely sampled co-infection time-series data. Moreover, the model accounts for seasonality in host reproduction. We illustrate an approach using a dataset describing co-infections by several strains of bacteria circulating within a population of cotton rats (*Sigmodon hispidus*). *Bartonella* strains were clustered into three genetically close groups, between which the divergence is correspondent to the accepted level of separate bacterial species. The proposed approach revealed no cross-immunity between genetic clusters while limited cross-immunity might exist between subgroups within the clusters.

Keywords: *Bartonella*, conditional least squares, cross-immunity, SIR model

1. Introduction

Multi-strain models have been widely used in epidemiology (Gupta, Ferguson, and Anderson, 1998; Kamo and Sasaki, 2002; Abu-Raddad et al, 2005; Bianco et al, 2009; Minayev and Ferguson, 2009). Developing and using multi-strain models is a challenging procedure due to numerous parameters such as death rate, birth rate, force of infection, and transmission rate, which are commonly assumed to be strain specific.

One of the key concepts of these models is cross-immunity, which allows infection by one strain to induce partial/perfect protection against other strains. Gupta, Ferguson, and Anderson (1998) proposed a very general model accounting for the cross-immunity in a multi-strain system, based on which they studied the effects of cross-immunity on evolution of strain structure. Abu-Raddad and Ferguson (2005) investigated population dynamics of host-pathogen systems involving an arbitrary number of antigenically distinct strains whose interaction depends on the cross-immunity. Minayev and Ferguson (2009) studied multi-strain deterministic epidemic models in which cross-immunity varies with the genetic distance between strains. Kamo and Sasaki (2002) proposed a two-strain susceptible-infected-recovered (SIR) model with cross-immunity. These models, however, assume an equilibrium population size over time, i.e., equal, constant birth and death rates. These assumptions might be too strong since the host population might fluctuate dramatically between seasons, which may affect the force of infection (Davis et al, 2005). For example, hispid cotton rat populations usually have peak litter production occurring in late spring and in late summer-early fall (Cameron and Spencer, 1984).

In addition, these earlier works were restricted to simulation studies assuming known parameter values. Another issue is that, except for the model of Gupta, Ferguson, and Anderson (1998), the state variables of these SIR-based models are often expressed in terms of the densities of various categories of the hosts. In general, it is often difficult to estimate the number of susceptibles, infectives, and recovered subjects over time, which makes an application of such models to real data challenging. Instead, developing a model consisting of proportions of susceptibles, infectives, and recovered subjects may make data analysis more feasible. Furthermore, some of state variables may not be observable in practice, for example, only infected individuals may be identified. Thus, it is pertinent to develop appropriate statistical models with partially observed data.

In this paper, we propose a two-strain SIR model that extends the model of Gupta Ferguson and Anderson (1998) and that of Kamo and Sasaki (2002) by accounting for seasonality in host reproduction and non-constant death rate. Furthermore, the proposed model is applied to a real dataset monitoring on the prevalence of co-infections of several *Bartonella* strains in a natural population of cotton rat (*Sigmodon hispidus*). This dataset is referred to as the Bartonella data and is described in Section 2. Details of the proposed two-strain SIR model and the statistical methods are elaborated in Sections 3 and 4, respectively. Interpretation and discussion of the epidemiological significance of the analysis results are given in Section 5. A brief conclusion is given in Section 6.

2. Data description

The field data used for this analysis were collected from a longitudinal study that monitored the prevalence of bartonella infection in a wild cotton rat population near Social Circle, Walton

County, Georgia, USA, over a period of 17 months, from March, 1996 to July, 1997, except December 1996, yielding altogether 483 trapping records (Kosoy et al., 2004a; Kosoy et al., 2004b). Cotton rats were captured for two or three consecutive nights each month and blood samples were taken. First-time captured cotton rats were marked. Marked and sampled rats were released. Sixty four out of 483 trapped rats were found to have co-infections by two or three *Bartonella* strains. Based on the cluster analysis of the genetic sequences among the bacterial isolates obtained from cotton rats (*Sigmodon hispidus*) in Georgia, identified *Bartonella* strains were clustered into three genogroups based on the similarities of the *gltA* sequences: A, B, and C (similarity range 88.2% to 93.5%). The citrate synthase gene, *gltA*, is a popular and widely used target to distinguish between closely related *Bartonella* species and genotypes (Kosoy, Hayman, and Chan, 2012). Since Norman et al. (1995) proposed the use of a variable fragment of this gene to differentiate *Bartonella*-like isolates at the species level, most laboratories working with bartonella bacteria have successfully applied this genetic marker. Birtles and Raoult (1996) have also demonstrated that the gltA-derived phylogeny appears to be more useful than the phylogeny derived from 16Sr DNA sequence data for investigating the evolutionary relationships of *Bartonella* species.

The three genogroups were further classified into unique sequence strains A1–A5, B1–B5, and C1–C2 with the sequence similarity ranged from 96.2% to 99.7%; see Table 1 of Kosoy et al. (2004b). In June and July of 1996, four cotton rats gave birth in their traps. To avoid issues related to vertical transmission of bartonella infection from parent subjects to their children, we excluded 19 neonatal rats captured in June and July. Fig. 1 shows the time-series plots of (i) the monthly numbers of trapped cotton rats (bottom figure) and (ii) proportions of trapped cotton rats that were infected by each *Bartonella* strain (top figure), which shows that A1 was the dominant

strain and the prevalence of strain B was low. In this paper, we consider two scenarios of cross-immunity: i) between genogroups; ii) between variants in the same genogroup. For the first scenario, we combined B and C due to low frequency of strain B and relatively high genetic similarity between strains B and C (Table 1 of Kosoy et al., 2004b). For the second scenario, we consider genogroup A only because of its high prevalence. Table 1 of Kosoy et al. (2004b) shows that A1 and A5 are genetically close to each other, and so are A2 and A4. Therefore, in this report, we compare A1&A5 vs. A2&A4 and A vs. B&C.

3. A two-strain SIR model with state variables expressed as proportions

We consider the two-strain special case of the multi-strain model proposed by Gupta, Ferguson, and Anderson (1998). Their model provides a general framework for modeling the dynamics of an infectious disease with multiple strains of a pathogen that may induce various degrees of cross-immunity in the hosts. Here, we extend their model to allow for variable host reproduction, and that the death rate can also be variable and not equal to the birth rate. Moreover, we modify the model so that a host is assumed to only make a fixed number of contacts with other hosts, on average. Detailed derivation of the model is given in Appendix.

The five state variables are: $x = x_{SS}$, $y_1 = x_{I\bullet}$, $y_2 = x_{\bullet I}$, $z_1 = x_{IS} + x_{RS}$, $z_2 = x_{SI} + x_{SR}$, where all variables are proportions of hosts with particular disease status indicated by the double subscripts with the first subscript being $S, I, R,$ standing for susceptible to the first strain, infected by the first strain, recovered from an infection by the first strain, and a subscript is set to $\bullet$ if no condition is imposed on the particular strain; the second subscript refers to the disease status with

respect to the second strain. For example, x_{IS} is the proportion of hosts infected by the first strain but susceptible to the second strain, $x_{I\bullet}$ is the proportion of hosts infected by the first strain, while $x_{\bullet I}$ is the proportion of hosts infected by the second strain. All state variables are implicit functions of time t with their derivatives denoted by the dot notation. The extended two-strain SIR model is given as follows:

$$\begin{aligned}
\dot{x} &= -\alpha_1 x y_1 - \alpha_2 x y_2 + (1-x)b, \\
\dot{y}_1 &= \alpha_1 (x + \delta z_2) y_1 - (\gamma_1 + b) y_1, \\
\dot{y}_2 &= \alpha_2 (x + \delta z_1) y_2 - (\gamma_2 + b) y_2, \\
\dot{z}_1 &= \alpha_1 x y_1 - \alpha_2 \delta z_1 y_2 - b z_1, \\
\dot{z}_2 &= \alpha_2 x y_2 - \alpha_1 \delta z_2 y_1 - b z_2,
\end{aligned} \tag{1}$$

where the parameter α_i's are the transmission rates between an individual infected by strain *i* (*i*=1, 2) and one susceptible to both strains, γ_i's are the host's recovery rate from an infection by strain *i*, δ is the cross-immunity parameter, and $b = b_t$ is the birth rate. Note that the death rate $\mu = \mu_t$ is eliminated in the algebra so that it no longer appears in (1). In other words, the death rate does not affect the dynamics when the state variables are expressed as proportions, i.e. $(x, y_1, y_2, z_1, z_2)^T$ in this model. Thus, (1) makes it feasible to analyze the general two-strain system without the need to know or to estimate μ. The non-negative parameter δ controls the degree of transmission of one strain of pathogen to hosts that have recovered from an infection or is infected by the other strain of the pathogen. In cases where a host acquires cross-immunity after recovery from an infection by one of the two strains, δ is less than 1 because of the reduction in transmission rate due to (partial) cross-immunity. Kamo and Sasaki (2002) and Gupta, Ferguson, and Anderson (1998) assumed that δ is between 0 and 1. However, in cases

where due to a weakened immune system by on-going infection, a host infected by one strain and susceptible to the other strain may have an elevated chance of being infected by the latter strain compared to a host susceptible to both strains (Small et al., 2010). For such cases, δ may be greater than 1. Thus, we shall allow δ to be non-negative. In summary, $\delta = 0$ represents the case of perfect cross-immunity between the two strains. If δ is positive and less than 1, there exists a partial cross-immunity between the two strains. If δ is equal to 1, there is no cross-immunity for the two strains and they infect the host independently. For $\delta > 1$, it signifies that the two strains are positively correlated, i.e., infection by one strain elevates the transmission rate of the other strain to the host. Asymmetric cross-immunity (Nuño et al., 2008) may be incorporated into the above model. However, in view of the relatively shortness of the *Bartonella* data, a parsimonious model may be desirable. Hence, in all model fitting reported below, we focus on the two-strain SIR model with symmetric cross-immunity and identical recovery rates.

4. Method

The state vector of the model defined by equation (1) is 5-dimensional. However, in practice, the state vectors may only be partially observed, i.e., only part of the state vector may be observable. For example, an observed host can be determined to be infected or non-infected with specific pathogen based on the laboratory tests, but for non-infected hosts, the question of whether these animals are susceptible or recovered may remain unknown, as this was the case for the *Bartonella* data (Kosoy et al., 2004a; Kosoy et al., 2004b). In other words, only the y_1 and y_2 component of the 5-dimensional state vector of equation (1) are observable for the bartonella data.

The main scientific question we address here concerns how similarity between two bacterial strains as deduced from their genetic sequences may relate to the host's cross-immunity to the strains. We explore this issue by estimating the host's cross-immunity against two pathogen groups of strains using the dataset discussed in Section 2. We first fit the proposed two-strain SIR model (1) to the monthly Bartonella infection rates by A1&A5 vs. A2&A4, where subgroups are combined due to their low prevalence and their relatively close genetic similarity. We then repeat the analysis contrasting A vs. B&C, with B and C merged into a group for a similar reason. In each analysis, the monthly observations consist of proportions of caught hosts infected by each of the two strains under study. Specifically, let the infection rate of the sampled hosts in the t^{th} month by strain i be denoted by $\tilde{y}_{i,t}, t = 1, 2$. These observed infection rates differ from the population infection rates $y_{i,t}$ by an additive measurement error: $\tilde{y}_{i,t} = y_{i,t} + \varepsilon_{i,t}$, where $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t})^T$ are independent random variables of zero mean and they are independent of the *y*'s. Let n_t be the number of hosts caught at time *t*. Then $(n_t \tilde{y}_{i,t}, i = 1, 2)$ has a trinomial distribution, so the variances of $\varepsilon_{i,t}$ are $\tilde{y}_{i,t}(1 - \tilde{y}_{i,t})$ and their covariance equal to $-\tilde{y}_{1,t}\tilde{y}_{2,t}$. To simplify model estimation, the distribution of ε_t is approximated by a bivariate normal distribution of zero mean and covariance matrix obtained with the unknown population infection rates replaced by the observed infection rates. It is an interesting future research problem of modifying the estimation scheme, to be elaborated below, that uses the exact sampling distribution of ε_t.

To allow for seasonal host reproduction, *b* is parameterized as a periodic function of a 12 month period:

$$b = b_t = u\sin(\frac{\pi}{6}t) + v\cos(\frac{\pi}{6}t) + w,$$

where *u, v* and *w* are unknown parameters.

The method of (approximate) conditional least squares via the unscented Kalman filter (UKF-CLS) (Ahn and Chan, 2013a) was employed to analyze the differential equation model (1) with the Bartonella data, which we now briefly outline. Consider the case that the state vector of the underlying system evolves according to a vector differential equation, with observations of some function of the state vector taken over discrete time. In our model, $\tilde{y}_t = (\tilde{y}_{1,t}, \tilde{y}_{2,t})^T$ is the observation vector and $v_t = (x_t, y_{1,t}, y_{2,t}, z_{1,t}, z_{2,t})^T$ the true state vector at time *t*. (For ease of exposition, we assume data were taken over equally-spaced epochs, say, $t = 1, 2, ..., n,$ but the method can be readily extended to irregularly sampled data.) Were the differential equation (1) linear and assuming normally distributed measurement errors, Kalman filter (Kalman, 1960) can be used to efficiently calculate the conditional mean $E_\theta(\tilde{y}_{(t|t-1)})$ and variance $Var_\theta(\tilde{y}_{(t|t-1)})$ of the predictive distribution of $\tilde{y}_t$ given past observations $\tilde{y}_s, s = 1, ...t-1,$ assuming model parameter θ. Note that the initial state vector v_0 is generally unknown, so it is included as part of parameter vector θ. However, Kalman filter is not applicable to a nonlinear system, e.g., the proposed two-strain SIR model. The unscented Kalman filter (UKF) proposed by Julier and Uhlmann (1997) is an iterative sampling-based algorithm that approximately computes the conditional means $\hat{y}_{t|t-1}(\theta) \approx E_\theta(\tilde{y}_{(t|t-1)})$ as well as the conditional variances, for a nonlinear process. For linear Gaussian processes, the UKF is essentially identical to the Kalman filter as it produces exact conditional means and variances. Moreover, the UKF provides high-order

approximation of the conditional means and variances for nonlinear processes, see Ahn and Chan (2013b). Unknown parameters can be estimated via approximate conditional least squares by minimizing the objective function $\sum_{t=1}^{n} | \tilde{y}_t - \hat{y}_{t|t-1}(\theta) |^2$ with respect to the unknown parameter θ, where $|\cdot|$ denotes the Euclidean norm of the enclosed expression and $\hat{y}_{t|t-1}(\theta)$ is the approximate conditional mean of $\tilde{y}_t$, given past observations, computed via the UKF. Ahn and Chan (2013a) derived some large-sample properties including consistency and asymptotic normality of the UKF-CLS estimator of a nonlinear system whose state is driven by some (stochastic) differential equation.

Theoretical standard errors of the UKF-CLS estimator derived in Ahn and Chan (2013a) for large samples can be used to construct confidence intervals and hypothesis testing. However, for small samples, confidence intervals may be alternatively computed via bootstrap as follows: i) suppose n_t rats were trapped in the t^{th} month, draw a simple random sample of size n_t from these captured rats, i.e., independent sampling with replacement and equal probability for each rat, and let the proportions infected by the two strains in the bootstrap sample be $(\tilde{y}^*_{1,t}, \tilde{y}^*_{2,t})^T$; ii) fit the two-strain SIR model with the bootstrapped series $\{(\tilde{y}^*_{1,t}, \tilde{y}^*_{2,t})^T, t = 1, ..., n\}$ to obtain the bootstrap parameter estimate θ^*; iii) repeat steps i) and ii) B, say 1000, times; iv) compute the 2.5 and 97.5 percentiles for each component of the B θ^*s which are the limits of the 95% bootstrap confidence interval. An advantage of this bootstrap scheme is that it preserves the dependence between the two strains, but note that the proposed bootstrap method is applicable

only if information of individual trapped host is available, which is, fortunately, the case for the *Bartonella* data.

5. Results

Tables 1 and 2 summarize the fitting results of the proposed two-strain SIR model with symmetric cross-immunity and identical recovery rates to the infection time series data with A1&A5 vs. A2&A4 and those with A vs. B&C, respectively. All 95% confidence intervals are obtained by nonparametric bootstrap detailed at the end of Section 4. The parameter α can be interpreted as the average number of infected rats per month, hence the monthly infection rate of rats with A1&A5, A2&A4, A, and B&C are 1.931, 1.722, 2.087, and 1.421, respectively. The γ parameter is the monthly recovery rate so the monthly recovery rates of cotton rats infected by A1&A5/A2&A4, and A/ B&C are 7.9% and 7.3%, respectively. (These estimated rates are broadly consistent with the results from a Markov chain analysis, see Table 5 in Chan and Kosoy (2010).) The estimated cross-immunity (δ) between A1&A5 and A2&A4 is 0.132 and its 95% confidence interval does not include 1. Thus, A1&A5 and A2&A4 enjoy partial cross-immunity against each other. On the other hand, the estimated cross-immunity between A and B&C is 0.55 and its 95% confidence interval includes 1, which suggests that A and B&C do not have a cross-immunity. We follow Nuño et al. (2005) in defining the basic reproductive number $R_i = \alpha_i / (\gamma + \mu)$ for the *i*th strain, which is the average number of secondary infections caused by the introduction of that strain in a fully susceptible population. For simplicity, we assume constant death rate which is set to be 1/6, because the average life span of a cotton rat is six months (Clark, 1972). The estimated basic reproductive numbers are 7.860 (95% CI: 5.075 – 12.930), 7.009 (95% CI: 2.493 – 13.136), 8.7079 (95% CI: 5.922 – 14.625), and 5.929 (95% CI:

1.164 – 15.438) for the two-strain model A1&A5 vs. A2&A4, and that of A, vs. B&C, respectively. These results are consistent with the observations that *Bartonella* infections by these strains were endemic, with infections predominantly by strain A, in the cotton-rat population under study.

Note that the estimates of the birth rate parameters *u*, *v*, and *w* are similar in both models. The bottom left curves in Fig. 2 show the estimated birth rate curve. The curve suggests that birth rate attains the maximum in June and the minimum in December, which is consistent with the report by Rose (1986) reporting that all of the trapped female cotton rats were pregnant from March through July, but none were pregnant in November and December. Recall that $\hat{y}_{t|t-1}(\theta)$ is the approximate conditional mean of $\tilde{y}_t$, given past monthly observations, computed via the UKF. The fitted values in the t^{th} month are then given by $\hat{y}_{t|t-1}(\hat{\theta})$ where $\hat{\theta}$ is the UKF-CLS estimate; the two components of the vector of fitted values will be denoted by $\hat{y}_{i,t}, t = 1, 2.$ The fitted values (red X's) are joined by red solid lines in Fig. 2, superimposed with the 95% predictive bounds (blue dotted lines) of the infection rates in Fig. 2, which track the observed infected proportions (solid circles) well. The 95% predictive intervals are computed by the formula $\hat{y}_{i,t} \pm 1.96 \times s_{i,t}$ where $s_{i,t}$ is the square root of the corresponding diagonal element of $Var_{\hat{\theta}}(\tilde{y}_{(t|t-1)})$ which is computed via UKF.

The residuals are defined as $r_{i,t} = \tilde{y}_{i,t} - \hat{y}_{i,t}$, i.e., subtracting the conditional means (fitted values) from the observed values, and the residuals estimate the error terms $e_{i,t} = \tilde{y}_{i,t} - E_{\theta_0}(\tilde{y}_{(t|t-1)})$ where

θ_0 is the true parameter. By construction, the $e_{i,t}$'s are independent. So, the goodness of fit of the fitted models may be assessed by checking whether the residuals are approximately independent. Residual diagnostics may be further simplified by standardizing the residuals by normalizing them by the estimate of the conditional standard deviations computed via the UKF. We examine whether or not the (standardized) residuals are autocorrelated by checking the residual autocorrelation functions (ACF), while between-series dependence in the residuals can be examined by the cross-correlation function (CCF), which are plotted in Fig. 3. None of the residual autocorrelations are significant and so are all cross-correlations, except for one lag, suggesting that the standardized residuals are uncorrelated over time. The independence assumption of the errors can be further validated by the Ljung-Box test. The Ljung-Box test statistic is proportional to the sum of squared residual autocorrelations from lag 1 to lag *k*. We conducted the Ljung-Box test with each residual series for *k* from 1 to 12, and all p-values are greater than 0.05, see Fig. 3. In summary, we conclude that the standardized residuals are approximately white noise, indicating that the models fit the data well.

6. Conclusion

A two-strain SIR epidemiological model with cross-immunity is applied to actual pathogen-host data and describes the kinetics of these relations. The fitted model suggests that A1&A5 and A2&A4 enjoy partial cross-immunity, but A and B&C sustain no cross-immunity. It is interesting to note that the estimated cross-immunity structure is broadly consistent with the genetic similarity pattern in Table 1 of Kosoy et al. (2004b). That is, if two strains are genetically similar, they are also epidemiologically similar in that they induce some cross-immunity in the cotton rats, but if they are genetically less similar, then they induce less cross-immunity in the

cotton rats. Importantly, this model fits current taxonomic requirements since genogroups A and B and C are discriminated on the level of the genetic divergence that was accepted by definition of a species level within genus of *Bartonella*, whereas the discrimination between A1&A5 and A2&A4 subgroups is under species level (La Scola et al., 2003). This phenomenon was also studied based on frequency and Markov chain analysis in Chan and Kosoy (2010). This model suggests that genetical relatedness can serve as a proxy for immunological protection between closely related bacterial strains, the characteristic that commonly remained unknown for most investigations of natural microbial communities. Further investigating the relationship between cross-immunity and genetic similarity is an interesting future research problem.

Acknowledgements

This work was partially supported by the U.S. National Science Foundation (DMS-1021896). The authors would like to thank Dr. Jennifer Le-Rademacher and two anonymous reviewers for their helpful comments and suggestions, which significantly improved the manuscript and Mr. Franco Mendolia for his computational assistance.

References

1. Abu-Raddad, L. J., Ferguson, N. M. 2005. Characterizing the symmetric equilibrium of multi-strain host-pathogen systems in the presence of cross-immunity. J. of Math. Biol. 50, 531–558.
2. Ahn, K. W., Chan, K. S. 2013a. Approximate conditional least squares estimation of a nonlinear state-space model via unscented Kalman filter. To appear in Comput. Stat. Data An.

3. Ahn, K. W., Chan, K. S. 2013b. On the convergence rate of the unscented transformation. To appear in Ann. I. Stat. Math.
4. Bianco, S., Shaw, L. B., Schwartz, I. B. 2009. Epidemics with multistrain interactions: The interplay between cross immunity and antibody-dependent enhancement. Chaos 19: 043123.
5. Cameron, G. N., Spencer, S. R. 1984. Sigmodon hispidus. Mammalian Species 158, 1–9.
6. Chan, K. S., Kosoy, M. 2010. Analysis of multi-strain Bartonella pathogens in natural host population – Do they behave as species or minor genetic variants? Epidemics 2, 165-172.
7. Clark, D. O. 1972. The extending of cotton rat range in California – Their life history and control. Proceedings of the 5th Vertebrate Pest Conference.
8. Davis, S, Calvet, E, Leirs, H. 2005. Fluctuating Rodent Populations and Risk to Humans from Rodent-Borne Zoonoses. Vector Borne Zoonotic Dis. 5, 305-314.
9. Gupta, S., Ferguson, N., Anderson, R. M. 1998. Chaos, persistence and the evolution of strain structure in populations of antigenically variable infectious agents. Science 240, 912-915.
10. Julier, S. J., Uhlmann, J. K. 1977. A new extension of the Kalman filter to nonlinear systems. Proc. of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando, FL Vol. Multi Sensor Fusion, Tracking and Resource Management II.
11. Kamo, M., Sasaki, A. 2002. The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model. Physica D 165, 228–241.

12. Kalman, R. E. 1960. New approach to linear filtering and prediction problems. J. Basic Eng-T ASMA D. 82, 35-45.
13. Kosoy, M., Hayman, D. T. S., Chan, K. S. 2012. Bartonella bacteria in nature: Where does population variability end a species start? Infect. Genet. Evol. 12, 894-904.
14. Kosoy, M., Mandel, E., Green, D., Marston, E., Jones, D., Childs, J. 2004a. Prospective studies of Bartonella of rodents. Part I. Demographic and temporal patterns in population dynamics. Vector Borne Zoonotic Dis. 4, 285–295.
15. Kosoy, M., Mandel, E., Green, D., Marston, E., Jones, D., Childs, J. 2004b. Prospective studies of Bartonella of rodents. Part II. Diverse infections in a single community. Vector Borne Zoonotic Dis. 4, 296–305.
16. Kosoy, M., Regnery, R.L., Kosaya, O.I., Childs, J.E., 1999. Experimental infection of cotton rats with three naturally occurring Bartonella species. J. Wildlife Dis. 35, 275–284.
17. La Scola, B., Zeaiter, Z., Khamis, A., Raoult, D. 2003. Gene-sequence-based criteria for species definition in bacteriology: the Bartonella paradigm. Trends Microbiol. 11, 318-321.
18. Minayev, P., Ferguson, N. M. 2009. Improving the realism of deterministic multi-strain models: implications for modeling influenza A. J. R. Soc. Interface 6, 509–518.
19. Nuño, M, Castillo-Chavez, C., Feng, Z., Martcheva, M. 2008. Mathematical models of influenza: the role of cross-immunity, quarantine and age-structure, Lecture Notes in Mathematics, vol. 1945, 349–361.

20. Nuño, M, Feng, Z., Martcheva, M., and Castillo-Chavez, C. 2005. Dynamics of two-strain influenza with isolation and partial cross-immunity, SIAM J. Appl. Math. 65, 964-982.

21. Rose, R. K. 1986. Reproductive strategies of meadow voles, hispid cotton rats, and eastern harvest mice in Virginia. Va. J. Sci. 37, 230-239.

22. Small, C. L., Shaler, C. R., McCormick, S., Jeyanathan, M., Damjanovic, D., Brown, E. G., Arck, P, Jordana, M., Kaushic, C., Ashkar, A. A., Xing, Z. 2010. Influenza infection leads to increased susceptibility to subsequent bacterial superinfection by impairing NK cell responses in the lung. J. Immunol. 184, 2048-2056.

Tables

Table 1. Parameter estimates of model (3) fitted to A1&A5 vs. A2&A4

	α_1	α_2	γ	δ
Estimates	1.931	1.722	0.079	0.132
95% CI	(1.683, 3.562)	(1.071, 3.686)	(0.043, 0.308)	(0.006, 0.437)
	u	v	w	
Estimates	0.345	0.136	0.781	
95% CI	(-0.121, 0.912)	(-0.483, 0.245)	(0.205, 1.238)	

Table 2. Parameter estimates of model (3) fitted to A vs. B&C

	α_1	α_2	γ	δ
Estimates	2.087	1.421	0.073	0.550
95% CI	(1.435, 3.687)	(0.197, 3.251)	(0.019, 0.172)	(0.012, 1.002)
	u	v	w	
Estimates	0.361	0.217	0.778	
95% CI	(0.126, 1.024)	(-0.427, 0.276)	(0.157, 1.020)	

Figure 1. Prevalence of *Bartonella* strains and the number of trapped cotton rats over the study period. Solid circles show observed values. The y-axis of the bottom figure represents the number of trapped cotton rats.

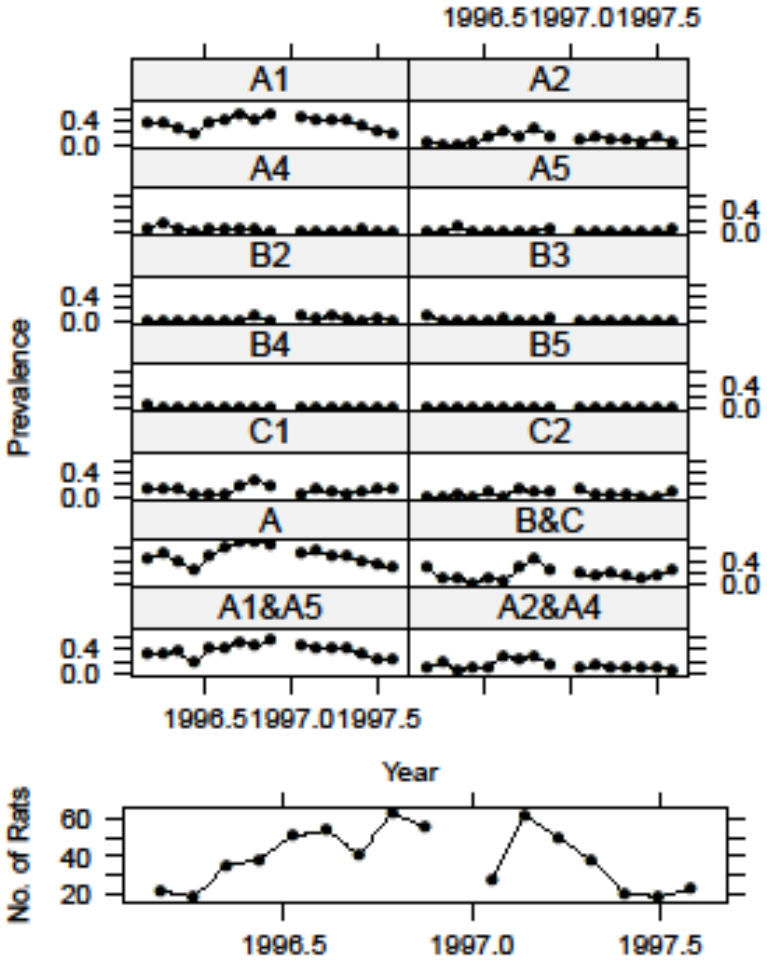


Figure 2. Estimated birth rate, prevalence and 95% confidence intervals. For the birth rate curve, the solid line is the birth rate function from the fitted model with A1&A5 vs. A2&A4 and the dotted line from that with A vs. B&C. The two estimated curves are quite similar. For the prevalence curves, the black dots and the red solid lines represent the observed and predicted infection rates, respectively. The blue dotted lines are the 95% confidence intervals.

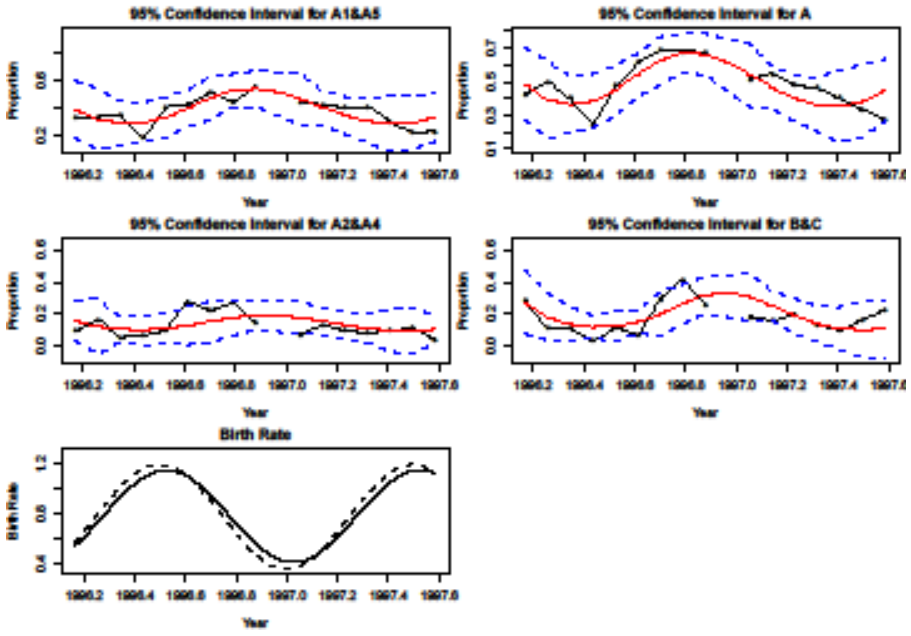
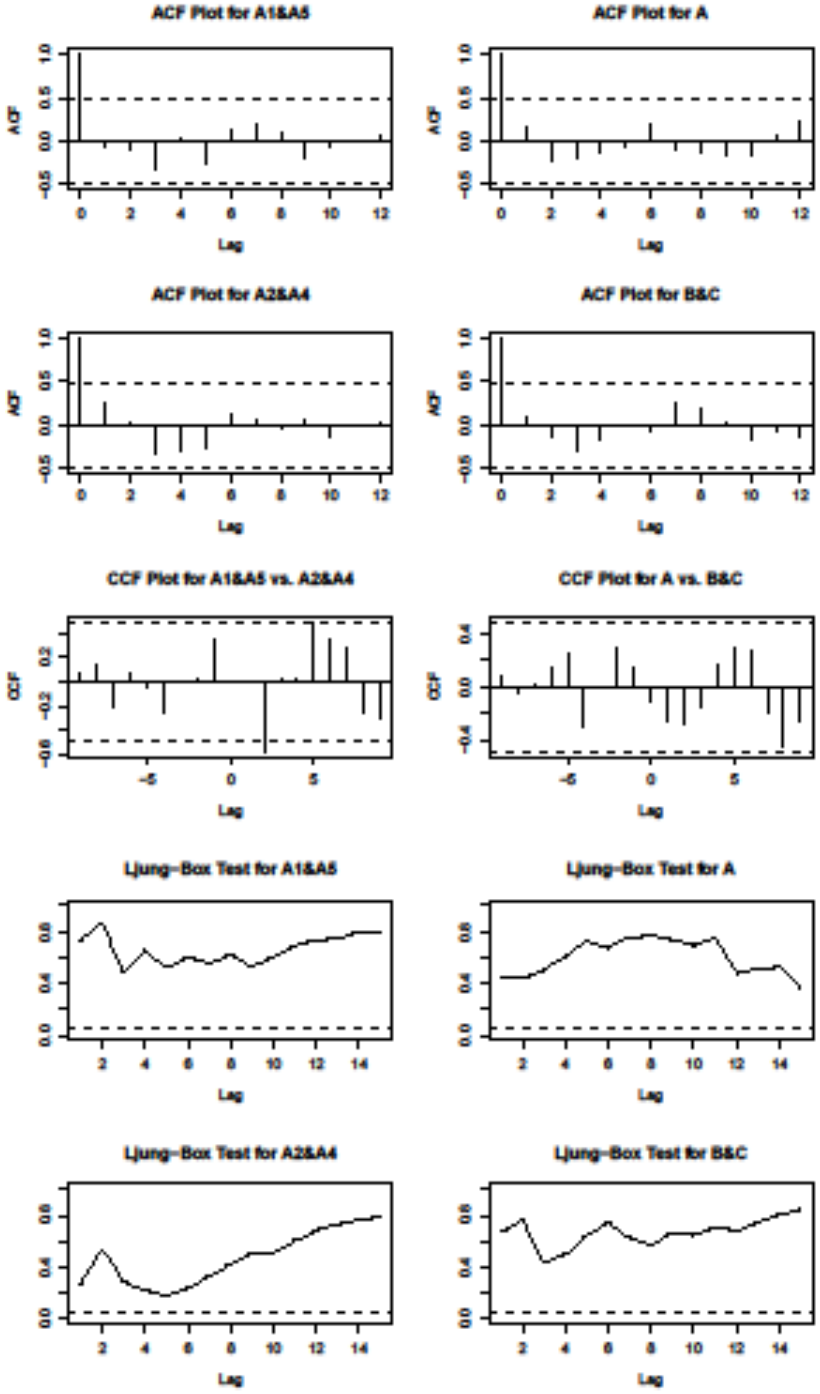


Figure 3. Diagnostics: ACF, CCF, and Ljung-Box p-value plots. All estimated residual autocorrelations lie within the (individual) 95% confidence intervals (dotted lines), suggesting that the residuals are not auto-correlated. The estimated residual cross-correlations are also within the 95% confidence intervals, indicating no cross-correlations in the residuals. All p-values in Ljung-Box plots are greater than 0.05 (dotted line), which further confirms that the residuals appear to be white noise.



Appendix
Derivation of Eqn. (1)

The multi-strain model proposed by Gupta, Ferguson, and Anderson (1998) provides a general framework for modeling the dynamics of an infectious disease with multiple strains of a pathogen that may induce various degrees of cross-immunity in the hosts. They model the state variables in terms of proportions but assume constant birth rate that is identical to the death rate. Kamo and Sasaki (2002) proposed a two-strain SIR differential equation model that admits cross immunity. They also assume constant death rate that is identical to the birth rate. However, their model has a more elaborate state vector (being 9-dimenional versus 5-dimensional for the two-strain special case of the model of Gupta, Ferguson and Anderson (1998)) with the state variables in terms of density. While there is operational advantage for having the state vector in terms of proportions, it is not clear how to incorporate variable birth and death rates in the model of Gupta, Ferguson and Anderson (1998). On the other hand, it is straightforward to incorporate variable birth and death rate in the model of Kamo and Sasaki (2002). Hence, we use the extended Kamo and Sasaki model to deduce the model for the disease dynamics for a two-strain system that admits cross immunity and variable birth and death rate, while the state variables are expressed in terms of proportions.

To formulate the proposed model, we first introduce some notations. Consider a host population subject to infection by two strains of the same pathogenic species. Each subject in the population may be classified as infected, recovered or susceptible according to its infection status concerning the first strain of pathogen and similarly classified regarding the second strain, resulting in nine categories *SS, IS, RS, SI, II, RI, SR, IR* and *RR*, where, e.g. *IS* stands for a subject infected by the first strain, but susceptible to the second strain. The densities of the nine

categories are denoted by X_{SS}, X_{IS}, etc., all of which are functions of time *t*, although the argument *t* is generally suppressed for simplicity. Their corresponding rates of change are denoted by $\dot{X}_{SS}$, $\dot{X}_{IS}$, etc. The Kamo-Sasaki model assumes constant population size, so the birth rate matches the death rate, which is often not true in reality. Thus, we modify the two-strain model of Kamo and Sasaki (2002) to include possibly variable birth and death rates as follows:

$$
\begin{aligned}
\dot{X}_{SS} &= -\beta_1 X_{SS} X_{I\bullet} - \beta_2 X_{SS} X_{\bullet I} - \mu X_{SS} + bN, \\
\dot{X}_{IS} &= \beta_1 X_{SS} X_{I\bullet} - \beta_2 \delta X_{IS} X_{\bullet I} - (\gamma_1 + \mu) X_{IS}, \\
\dot{X}_{RS} &= \gamma_1 X_{IS} - \beta_2 \delta X_{RS} X_{\bullet I} - \mu X_{RS}, \\
\dot{X}_{SI} &= \beta_2 X_{SS} X_{\bullet I} - \beta_1 \delta X_{SI} X_{I\bullet} - (\gamma_2 + \mu) X_{SI}, \\
\dot{X}_{II} &= \beta_1 \delta X_{SI} X_{I\bullet} + \beta_2 \delta X_{IS} X_{\bullet I} - (\gamma_1 + \gamma_2 + \mu) X_{II}, \\
\dot{X}_{RI} &= \beta_2 \delta X_{RS} X_{\bullet I} + \gamma_1 X_{II} - (\gamma_2 + \mu) X_{RI}, \\
\dot{X}_{SR} &= \gamma_2 X_{SI} - \beta_1 \delta X_{SR} X_{I\bullet} - \mu X_{SR}, \\
\dot{X}_{IR} &= \beta_1 \delta X_{SR} X_{I\bullet} + \gamma_2 X_{II} - (\gamma_1 + \mu) X_{IR}, \\
\dot{X}_{RR} &= \gamma_1 X_{IR} + \gamma_2 X_{RI} - \mu X_{RR}, \\
\dot{N} &= bN - \mu N,
\end{aligned}
\tag{1}
$$

where *b* is the (possibly variable) birth rate, μ the (possibly variable) death rate and *N* is the total population size. Since we do not consider vertical transmission in this paper, we assume that newborns (*bN*) are susceptible to both strains, which is, e.g., codified in the first equation of (1). The parameters β_i and γ_i are the transmission rate and the recovery rate for the strain *i*, *i*=1,2. Note that the β_i's are the transmission rates under the assumption that a host may make random contacts with any host in the population. The parameters δ and μ represent the degree of cross-immunity and the death rate, respectively.

Kamo and Sasaki (2002) assumed that δ varies between 0 and 1. For $\delta = 0$, a host recovered from an infection by one strain of the pathogen acquires perfect cross-immunity against other strain, that is, it will never be infected by the other strain. On the other hand, for $\delta = 1$, a host recovered from an infection by one strain of the pathogen does not have a cross-immunity against the other strain, so that the two strains can independently infect a host. One of the assumptions for the two-strain SIR model (1) is that a host may make contact with any host in the population and that the contact rate is proportional to the population size, which may not hold for the case of vast study area and/or large population.

To derive the two-strain model with the state expressed in terms of proportions, we consider the first derivative of the proportions, for example, $\frac{d}{dt}(X_{SS} / N)$. The variables standing for the proportions will be denoted in lower case, e.g. we write x_{SS} for X_{SS} / N, etc. Following Kamo and Sasaki (2002), we transform the nine state variables into five state variables as follows: $x = x_{SS}$, $y_1 = x_{I\bullet}$, $y_2 = x_{\bullet I}$, $z_1 = x_{IS} + x_{RS}$, $z_2 = x_{SI} + x_{SR}$, where $x_{I\bullet} = x_{IS} + x_{II} + x_{IR}$ and $x_{\bullet I} = x_{SI} + x_{II} + x_{RI}$. In addition, we replace the term $\beta_1 X_{SS} X_{1\bullet}$ by $\alpha_1 X_{SS} X_{1\bullet} / N$ where α_1 is the product of the expected number of contacts a host makes with other hosts per unit time and the transmission probability given a contact between an individual infected by strain 1 and one susceptible to both strains. This specification implies that on average a host makes a fixed number of contacts per unit of time.

Next, we consider the first derivative of X_{SS}, $(X_{SS} / N)'$. Then, we have

$$\begin{aligned}\left(\frac{X_{SS}}{N}\right)' &= \frac{\dot{X}_{SS}N - X_{SS}\dot{N}}{N^2} = \frac{\dot{X}_{SS}}{N} - \frac{X_{SS}\dot{N}}{N^2} \\ &= \frac{-\beta_1 X_{SS}X_{I\bullet} - \beta_2 X_{SS}X_{\bullet I} - \mu X_{SS} + bN}{N} - \frac{X_{SS}(bN - \mu N)}{N^2} \\ &= \frac{-\beta_1 X_{SS}X_{I\bullet} - \beta_2 X_{SS}X_{\bullet I} - bX_{SS} + bN}{N} \\ &= -\alpha_1 x_{SS}x_{I\bullet} - \alpha_2 x_{SS}x_{\bullet I} - (1 - x_{SS})b.\end{aligned}$$

Note that μ was eliminated. Similarly, we can derive the following:

$$\left(\frac{X_{IS}}{N}\right)' = \alpha_1 x_{SS}x_{I\bullet} - \alpha_2 \delta x_{IS}x_{\bullet I} - (\gamma_1 + b)x_{IS},$$

$$\left(\frac{X_{RS}}{N}\right)' = \gamma_1 x_{IS} - \alpha_2 \delta x_{RS}x_{\bullet I} - bx_{RS},$$

$$\left(\frac{X_{SI}}{N}\right)' = \alpha_2 x_{SS}x_{\bullet I} - \alpha_1 \delta x_{SI}x_{I\bullet} - (\gamma_2 + b)x_{SI},$$

$$\left(\frac{X_{II}}{N}\right)' = \alpha_1 \delta x_{SI}x_{I\bullet} + \alpha_2 \delta x_{IS}x_{\bullet I} - (\gamma_1 + \gamma_2 + b)x_{II},$$

$$\left(\frac{X_{RI}}{N}\right)' = \alpha_2 \delta x_{RS}x_{\bullet I} + \gamma_1 x_{II} - (\gamma_2 + b)x_{RI},$$

$$\left(\frac{X_{SR}}{N}\right)' = \gamma_2 x_{SI} - \alpha_1 \delta x_{SR}x_{I\bullet} - bx_{SR},$$

$$\left(\frac{X_{IR}}{N}\right)' = \alpha_1 \delta x_{SR}x_{I\bullet} + \gamma_2 x_{II} - (\gamma_1 + b)x_{IR},$$

$$\left(\frac{X_{RR}}{N}\right)' = \gamma_1 x_{IR} + \gamma_2 x_{RI} - bx_{RR},$$

$$\left(\frac{N}{N}\right)' = 1' = 0.$$

After transforming $x = x_{SS}$, $y_1 = x_{I\bullet}$, $y_2 = x_{\bullet I}$, $z_1 = x_{IS} + x_{RS}$, and $z_2 = x_{SI} + x_{SR}$, we obtain the following system of equation which is identical to Eqn. (1) in the main text:

$$
\begin{aligned}
\dot{x} &= -\alpha_1 x y_1 - \alpha_2 x y_2 + (1-x)b, \\
\dot{y}_1 &= \alpha_1 (x + \delta z_2) y_1 - (\gamma_1 + b) y_1, \\
\dot{y}_2 &= \alpha_2 (x + \delta z_1) y_2 - (\gamma_2 + b) y_2, \\
\dot{z}_1 &= \alpha_1 x y_1 - \alpha_2 \delta z_1 y_2 - b z_1, \\
\dot{z}_2 &= \alpha_2 x y_2 - \alpha_1 \delta z_2 y_1 - b z_2,
\end{aligned}
\tag{2}
$$

where the parameter α_i's are the transmission rates between an individual infected by strain *i* (*i*=1, 2) and one susceptible to both strains. Note that the death rate parameter μ is eliminated in the algebra so that it no longer appears in (2). However, since Equation (2) is directly derived from (1), (2) accounts for the possibility that the birth rate differs from the death rate. In other words, the death rate does not affect the dynamics when the state variables are expressed as proportions, i.e. $(x, y_1, y_2, z_1, z_2)^T$ in this model.