University of New Mexico
UNM Digital Repository

Biology ETDs | Electronic Theses and Dissertations

7-1-2014

Population and Metapopulation Ecology of Childhood Diseases

Christian E. Gunning

Follow this and additional works at: https://digitalrepository.unm.edu/biol_etds

Recommended Citation
Gunning, Christian E.. "Population and Metapopulation Ecology of Childhood Diseases." (2014). https://digitalrepository.unm.edu/biol_etds/46

This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at UNM Digital Repository. It has been accepted for inclusion in Biology ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.

Christian E. Gunning

Candidate

Biology

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Helen J. Wearing , Chairperson

James H. Brown

Erik B. Erhardt

Melanie E. Moses

Population and Metapopulation Ecology of Childhood Diseases

by

Christian E. Gunning

B.S., University of Georgia, 2001
M.W.R., Hydroecology, University of New Mexico, 2009

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biology

The University of New Mexico

Albuquerque, New Mexico

July, 2014

©2014, Christian E. Gunning

Dedication

To Albuquerque, where I spent this decade of wonder and exploration. To this curious urban playground, now covered in memories: lavender & hollyhocks & datura; ash & mesquite & juniper; sunflowers & hummingbirds; agaves & crows; arroyos of sand and clay and concrete; countless paths on foot & bike; low-hanging apricots, plums, & figs; from spring dust storms to monsoon lightning to the autumnal smells of roasting green chile; and the most beautiful sunrises, of which I've seen plenty.

Acknowledgments

I have received a surfeit of very good advice from Dr. Helen Wearing, without which I would have long ago given up; from Dr. Jim Brown, who encouraged me to enjoy science and have fun; and from Dr. Erik Erhardt, who patiently walked me through this-that-and-the-other model. My late-night confidants, for their humor and insight and all-around indulgence: Robert Liberatore, Nichole Y. Carnevale, Monet Maloof, Lenore Gulsch, Nicolas Giron, and Michael Chang. I am endlessly appreciative of my family's patience with my wayward progress into adulthood. Finally, I am deeply grateful for the unwavering support of Natalie Wright, the blue-eyed blessing.

Population and Metapopulation Ecology of Childhood Diseases

by

Christian E. Gunning

B.S., University of Georgia, 2001
M.W.R., Hydroecology, University of New Mexico, 2009
Ph.D, Biology, University of New Mexico, 2014

Abstract

Researchers have long used mathematical models and empirical data to explore the population ecology of childhood diseases such as measles and whooping cough. These diseases have proven ideal model systems for studying population dynamics over space and time. Here we present a novel dataset of weekly measles and whooping cough case reports in pre-vaccine era U.S. cities and states, along with a previously-studied dataset of measles in England & Wales.

We first estimate per-population disease reporting probabilities. We find that disease reporting is highly variable over space and between diseases, and correlated with socioeconomic covariates including ethnic composition and school attendance. Using these reporting estimates, we infer the long-term, marginal distribution of disease incidence for each population. This describes a probabilistic measure of disease persistence that compares favorably with a classic threshold persistence measure, *critical community size* (CCS). The U.S. and England & Wales exhibit similar patterns of measles incidence distributions: larger populations show higher mean

incidence and lower variance. The per-time probability of local extinction (conditioned on population size) is higher in the U.S. than in England & Wales, likely due to larger distances between U.S. cities. Finally, we use observed persistence and inferred incidence distributions to estimate the per-time probability of true persistence. Estimated persistence of whooping cough is much higher than persistence of measles (conditioned on population size). We find that cryptic persistence (the difference between observed and estimated persistence) of whooping cough is most common in small populations, while for measles cryptic persistence is most common in medium-sized populations that hover at the edge of extinction.

Our results show that variation in disease reporting can significantly affect metapopulation estimates of disease persistence, such as CCS. The distributional estimates of incidence presented here explicitly account for incomplete reporting, providing summaries of long-term ecological patterns that are comparable between metapopulations. These measures can provide disease control programs with valuable information on where disease incidence is expected to be higher or lower than expected based on population size alone.

Contents

Chapter 1

Introduction

Childhood disease is the name given to any of a cluster of diseases that are all directly-transmitted, acutely infectious, and fully immunizing. These diseases, including measles, pertussis, diptheria, mumps, and chicken pox, infected the vast majority of individuals in the pre-vaccine era at an early age. In the modern era, vaccination for these diseases is almost universal in developed nations, and overall incidence is very low.

Their relative dynamical simplicity (early age of infection, no repeated infection) coupled with dramatic epidemic cycles and high-fidelity historical records of disease incidence have made childhood diseases useful model systems in population and disease ecology [1–4]. An historical overview of this extensive body of literature is given in the following chapters. Suffice to say that this dissertation represents the tip of a long line of disease ecology research that couples mathematical and statistical models with empirical observations of childhood diseases, particularly measles and pertussis.

The key question driving this research is to identify the necessary and sufficient conditions for these diseases to persist. Large population sizes, high host birth rates,

and spatial connectivity are all well-known drivers of disease persistence. Pathogen life history traits, including basic reproductive ratio (R_0) and infectious period, shape patterns of disease persistence. A key goal in disease ecology is to disentangle the effects of local dynamics, metapopulation processes, and the innate characteristics of pathogens on disease persistence.

Much of the research to-date has focused on patterns of disease incidence in England & Wales, largely due to the availability of high-quality data sources. Spatial connectivity and isolation are hypothesized to affect local and metapopulation disease dynamics, and the applicability of findings based on this densely-settled island nation to other systems warrants testing. Indeed, a primary goal of the research presented here was to test findings from England & Wales in a qualitatively different metapopulation. By comparison, the U.S. is geographically larger, less densely populated, and more ethnically diverse than England & Wales.

This particular body of work emerged as an extension to Bartlett's classic threshold measure of disease persistence, critical community size (CCS) [5]. According to classical epidemiological theory, the metapopulation persistence of childhood diseases like measles is driven by cities above a critical size, the CCS, where susceptible individuals are replenished quickly enough so that a local chain of infection proceeds unbroken. While competing definitions of CCS have been employed, none have accounted for incomplete and variable disease reporting.

In Chapter 2 [6], we provide a comprehensive, probabilistic interpretation of CCS that we apply to case reports of measles in cities in the U.S. and England & Wales. We find that both metapopulations exhibit similar patterns of measles incidence distributions: larger populations show higher mean incidence and lower variance. On the other hand, the per-time probability of local extinction (conditioned on population size) is higher in the U.S. than in England & Wales, likely due to larger distances between U.S. cities.

In Chapter 3 [7], we provide a comprehensive review of disease reporting variability for both measles and whooping cough in the U.S. and England & Wales. We refine the methods employed in Chapter 2, as well as develop uncertainty bounds on estimates of incomplete reporting for diseases in the U.S. We also find that, in the U.S., disease reporting correlates with socioeconomic covariates including ethnic composition and school attendance.

Finally, Chapter 3 [8] employs empirical patterns of observed persistence and inferred incidence distributions to estimate the per-time probability of true persistence. This allows us to disentangle the effects of poor disease reporting on observed case reports from the effects of disease extinction on patterns of incidence. We define cryptic persistence as the difference between observed and estimated persistence, and show that patterns of cryptic persistence differ markedly between diseases.

This body of work contributes to disease and population ecology by providing tools to more accurately assess patterns of true incidence over space and time. In addition, this work is applicable to modern disease control efforts, potentially aiding epidemiologists and public health professionals identify hotspots of disease incidence, as well as populations at greatest risk of cryptic persistence.

1.1 References

[1] WH Hamer. The Milroy lectures on epidemic disease in England – the evidence of variability and persistence of type. *Lancet*, 1:733–739, 1906.

[2] H.E. Soper. The interpretation of periodicity in disease prevalence. *J R Stat Soc*, 92(1):34–73, 1929.

[3] M.S. Bartlett. *Stochastic population models in ecology and epidemiology*. Methuen London, 1970.

[4] R.M. Anderson and R.M. May. *Infectious diseases of humans*. Oxford University Press Oxford, 1991.

[5] M.S. Bartlett. The critical community size for measles in the United States. *J R Stat Soc Ser A*, 123(1):37–44, 1960.

[6] C.E. Gunning and H.J. Wearing. Probabilistic measures of persistence and extinction in measles (meta)populations. *Ecol. Lett.*, 16:985–994, 2013.

[7] C.E. Gunning, E. Erhardt, and H.J. Wearing. Incomplete reporting of pre-vaccine era childhood diseases: a case study of observation process variability. *Proc. R. Soc. B*, In Review.

[8] C.E. Gunning, M. Ferrari, and H.J. Wearing. Cryptic persistence in childhood disease. In Prep.

Chapter 2

Inferring Distribution of Incidence

Title Probabilistic measures of persistence and extinction in measles (meta)populations

Christian E. Gunning[1], Helen J. Wearing[2]

1 Department of Biology, University of New Mexico, Albuquerque, NM, USA. Email: xian@unm.edu.

2 Department of Biology and Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, USA. Email: hwearing@unm.edu.

Citation: C.E. Gunning and H.J. Wearing. Probabilistic measures of persistence and extinction in measles (meta)populations. *Ecology Letters*, 16:985-994, 2013.

Author Contributions: CEG and HJW designed the study. CEG authored the data collection system, performed the analyses and wrote the first draft of the manuscript. HJW contributed substantially to revisions.

2.1 Abstract

Persistence and extinction are fundamental processes in ecological systems that are difficult to accurately measure due to stochasticity and incomplete observation. Moreover, these processes operate on multiple scales, from individual populations to metapopulations.

Here we examine an extensive new dataset of measles case reports and associated demographics in pre-vaccine era U.S. cities, alongside a classic England & Wales dataset. We first infer the per-population quasi-continuous distribution of log incidence. We then use stochastic, spatially implicit metapopulation models to explore the frequency of rescue events and apparent extinctions. We show that, unlike critical community size, the inferred distributions account for observational processes, allowing direct comparisons between metapopulations.

The inferred distributions scale with population size. We use these scalings to estimate extinction boundary probabilities. We compare these predictions with measurements in individual populations and random aggregates of populations, highlighting the importance of medium-sized populations in metapopulation persistence.

2.2 Introduction

The persistence (and extinction) of species over space and time is an emergent property of multiple ecological processes. From conservation biology to disease ecology, local population dynamics and spatial connectivity are central to understanding species persistence. Host-pathogen interactions provide a natural framework in which to consider colonizer-invader trade-offs and the prerequisites for successful invasion and persistence [1]. These problems have straightforward epidemiological and public health interpretations: the emergence and establishment of novel human pathogens,

and their control and eradication in human populations. More generally, the study of disease dynamics can shed light on the ecologically significant interactions between demographic stochasticity, patch connectedness, and observational processes.

Measles has been a workhorse of theoretical ecology for more than 100 years [2–7], particularly with respect to population-level persistence and extinction. A human pathogen that diverged from rinderpest in the pre-industrial era, measles has avoided eradication and still causes significant morbidity and mortality, particularly in regions with poor health care infrastructure [8]. Originally noted for its dramatic yet regular epidemics, measles has become a model system in population ecology modeling. Epidemiological models of measles highlight the importance of non-linear feedbacks, transients, stochasticity, and non-stationarities in ecological processes [9–12], as well as spatial structure and heterogeneity [12–14].

Several factors make measles a useful model pathogen, including ease of diagnosis, abundant historical records, short latent and infectious periods, low mortality, lifetime immunity, and a lack of environmental or animal reservoirs. Seasonal aggregation of school-age children in developed countries also influences the basic reproductive ratio (R_0), controlling epidemic timing [11, 15, 16]. These factors facilitate the interpretation of historical records by reducing uncertainty and constraining dynamics.

Population measures of persistence

Critical community size (CCS) is a threshold measure of within-population disease persistence. CCS has played a central organizing role in disease ecology since its introduction by Bartlett [17]. Though debate surrounds its precise definition [17–20], CCS is approximately the population size above which pathogen extinction is not observed, implying an unbroken within-population chain of infection. The CCS

of measles has been extensively studied in the context of pre-vaccine era England & Wales [21], and with respect to vaccination and eradication thresholds [22, 23].

As defined, CCS depends fundamentally on both population and metapopulation-level processes, and serves as a measure of both. This ambiguity of the CCS concept is highlighted by vaccination, which is dynamically equivalent to lowering birth rates. As susceptible recruitment drops, the CCS of a given metapopulation increases. In the extreme case, endemic transmission of measles was eliminated in the U.S. circa 2000 [24]. Using the traditional definition of CCS, the current CCS of the U.S. exceeds the size of every U.S. city, by definition, since continuous transmission is no longer observed in any U.S. city.

Minimum viable metapopulation (MVM) size is a complementary threshold measure of persistence that addresses this ambiguity, though does not consider individual population sizes [25]. The present work seeks, in part, to bridge a perceived gap between classic disease ecology and metapopulation literature, since these fields have historically approached similar questions from very different directions (e.g. the contemporaneous work of Hanski et al. [25] and Bolker and Grenfell [13]).

As measured, CCS is further confounded by observational processes such as sampling period and reporting rate, such that no straightforward between-metapopulation comparison exists. Here we present a new dataset consisting of over 20 years of weekly case reports of measles in 83 cities in the pre-vaccine era United States (from 1924 to 1945). Case reports are augmented by demographic records at the city, state and national level, and associated reporting rate estimates. We compare this dataset with the classic England & Wales dataset (biweekly, 1944 to 1965), also augmented with demographics and reporting rate estimates. We demonstrate the sensitivity of CCS to observational, within-population, and metapopulation processes, and provide an alternate measure of persistence.

We seek summary statistics that, similar to CCS, describe the long-term, marginal distribution of true measles incidence within and between populations. Here we use a distributional approach, augmented by stochastic metapopulation models that highlight the key dynamical and observational processes in these systems. Despite the presence of obvious autocorrelation and non-stationarity, we find conserved measures of measles incidence that scale with population size.

Three points deserve special notice. First, we find that U.S. reporting rates are lower and more variable than in England & Wales. This suggests a larger apparent CCS, for which we have corrected. Second, higher U.S. birth rates should favor increased persistence, which we do not observe. Finally, large U.S. intercity distances could result in smaller rescue effects and lead to decreased persistence, as observed here. These findings suggest that metapopulation processes play an important role in differentiating the U.S. from England & Wales. Nonetheless, a key remaining challenge in matching mechanistic models to data is the full differentiation of population-level processes such as seasonality of transmission from metapopulation-level processes such as migration.

Outline

An outline of the paper is as follows. We first infer weekly log incidence (ξ) from observed cases, reporting rate, and population size. Critically, we do not exclude observed zeros from our analysis. Instead we assume that each represents ≤ 1 observed case. We then fit a normal cumulative distribution function (CDF) to each city's empirical cumulative distribution function (ECDF). Scaling of the inferred parameters (mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$, not sample mean and SD) with population size (N) is evident. For each inferred parameter and metapopulation, we fit a descriptive linear model using N as the independent variable. We construct a probabilistic

measure of CCS, and compare patterns of persistence between random aggregates of populations and metapopulations.

For comparison, we construct a stochastic, spatially-implicit metapopulation model that includes fully-parametrized demographics. Model results highlight the effects of apparent extinctions on the proportion of observed zeros, particularly in intermediate-sized cities. We conclude with a discussion of the applicability and usefulness of the presented methodology to other systems.

2.3 Materials and Methods

Data collection and preparation

U.S weekly case reports were manually transcribed from United States Public Health Reports [26]. Each report was double-entered; mis-matches were automatically identified and manually resolved. Populations with fewer than 20% missing values were used for subsequent analysis, for a total of 83 cities ranging in mean population from 16 thousand to 7.2 million. These cities account for 22.7% (1920) to 24.9% (1950) of the total U.S. population, and from 48% (1920) to 39% (1950) of the urban U.S. population. All cities with a 1930 population over 350 thousand are sampled; many smaller cities are missing. The period of record stretches from 1924-01-05 to 1945-12-29 (1148 weeks).

U.S. city decadal population was obtained from the U.S. decadal census (1920-1950) [27]. State per capita birth, death, and infant mortality rates were obtained from the *U.S Statistical Abstracts, 1920-1950* [28]. Yearly U.S. city populations were estimated using an exponential growth model to interpolate between decadal population. Yearly U.S. city populations were then used to calculate births into each city from state birth and national infant mortality rates. Birth rates for the

U.S. (states) and England & Wales (cities) are shown in Figure 2.5 in Supporting Information.

England & Wales case reports (every two weeks, no missing values, as presented by Grenfell et al. [21]) were obtained from
`http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.htm`, along with population size [29] and births by year. Births were adjusted for infant mortality using yearly national rates [30]. This dataset includes 60 cities ranging in (1955) population from 10.5 thousand to 3.25 million. The period of record covers 1944-01-09 to 1966-12-25; due to redistricting changes in 1965, only the period through 1964 was employed, resulting in 546 biweeks.

For each population, migration was inferred by subtracting yearly births from the yearly change in population size. A proportion of migrants $(1 - 1/R_0)$ was assumed to be recovered, with the remainder susceptible.

Reporting Rate

Reporting rate was assumed constant over the period of record; for the i^{th} population, a single reporting rate r_i was estimated. We assume that a proportion $1 - \frac{1}{R_0}$ of available susceptibles contract and recover from measles [6] over the period of record. The net yearly flow of susceptibles s_i into population i was estimated from births, infant mortality, and migration, as described above. Death of susceptibles was assumed to be minimal. Thus, the expected total number of actual cases in the i^{th} population, $\overline{C_i} = \sum_t s_i(1 - \frac{1}{R_0})$, and $r_i = \frac{\sum_t C_i}{\overline{C_i}}$, where C_i represents the observed case reports in population i at time t [16]. See Figure 2.6 for a map of sampled U.S. cities showing estimated reporting rates.

Estimating lognormal distributions of incidence

For each city, we inferred the distribution of weekly per capita log incidence ξ_i ($\log_{10}$ was used throughout). Missing values were excluded. Inferred cases $\hat{C}_i$ were estimated from reported cases C_i and reporting rate: $\hat{C}_i = \frac{C_i}{r_i}$. Critically, we do not exclude observed zeros to avoid distorting the ECDF. Instead, we assume that each observed zero is equivalent to as many or fewer inferred cases as one observed case: $C_i = 0 \Rightarrow \hat{C}_i \leq \frac{Z}{r_i}$ for $Z \leq 1$ (here $Z = 1$).

Weekly per capita log incidence ξ_i was estimated from inferred cases, the mean population size of each city N_i, and the number of weeks per observation n (2 for England & Wales): $\xi = \log \frac{\hat{C}_i}{nN_i}$. Nonlinear minimization (NLM) [31] was used to fit a normal CDF to the ECDF of ξ_i using an L_∞ metric (equivalent to the Kolmogorov-Smirnoff (K-S) distance between the two distributions). Metapopulation mean $\hat{\mu}$ and $\hat{\sigma}$ were used as initial conditions for another iteration of the NLM procedure to avoid local minima. Thus, the mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ were chosen to minimize the maximum difference in probabilities (L_∞ metric) between each population's empirical and estimated CDF of ξ.

In this way, we infer a quasi-continuous distribution of log incidence ξ from the discretized distribution of observed cases. The ECDFs and estimated normal CDFs are shown for select cities in Figure 2.8. This inference method explicitly accounts for the proportion of observed zeros as the integral of the normal probability density function (PDF) over the interval $(-\infty, \log \frac{1}{rN_i})$. Conceptually, this lower tail includes inferred cases below the observation threshold of $\hat{C}_i = \frac{1}{r}$, as well as the effects of imported cases. To evaluate goodness-of-fit, parametric and non-parametric bootstrap replicates were conducted (see Figure 2.11).

To test for differences between metapopulations, we conducted an (unbalanced) ANCOVA using country identity as the independent variable and log population size

N as the covariate. Separate linear models were tested for $\hat{\mu}$ and $\hat{\sigma}$.

A metapopulation model of measles dynamics

We assessed the ability of simple epidemiological models to reproduce patterns observed in the data. We constructed a spatially implicit, stochastic, event-driven version of the standard exponential SEIR model as per Olsen et al. [32]. The resulting simulations also highlight unobservable yet important processes, including rescue events and apparent extinctions.

We employed the Gillespie τ-leap method [33] with a time-step of one day. Population sizes and demographics were fully specified from historical records. Births into the susceptible class account for infant mortality. We assume death occurs exclusively in the recovered class. A portion of migrants ($\frac{1}{R_0}$) was assumed to be susceptible; the remaining migrants enter or leave the population through the recovered class. Transitions into the exposed class due to imported infection were included at a rate proportional to metapopulation incidence. We tested both sinusoidal and term-time seasonal forcing of contact rates. Unlike England & Wales, U.S. term times are not national, and historical estimates are not available. The effect of varying term times remains under-explored in the literature. See Appendix S2 for model details.

Key transitions were summed on a weekly basis. These include total transitions into E (E_w), transitions into E caused by imports ($E_{w\eta}$), and total transitions into I (I_w). A binomial observation process was used to generate weekly observed cases I_{wo} from I_w, where the probability of successful observation was equal to the city's reporting rate r_i. For $\Gamma = E_w + I_w$, we tabulated the total number of true extinctions ($\Gamma = 0$), rescue events ($\Gamma = E_{w\eta}$), and apparent extinctions ($\Gamma > 0 \cap I_{wo} = 0$). The proportion of weeks with zero case reports ($P_0 = Pr(I_{wo} = 0)$) was also tabulated.

An ensemble of 10 realizations was simulated for each of a range of parameter

combinations (see Table 2.4). For each realization and population i, P_{0i}, $\hat{\mu}_i$, and $\hat{\sigma}_i$ were computed. For each of these 3 measures δ, the sum of squared residuals $RSS_\delta = \sum_i \delta_{i,model} - \delta_{i,data}$ was computed. See Figure 2.13 for final model selection details and Table 2.5 for final parameter values. For these parameter values, an ensemble of 50 realizations was run and within-city ensemble means of all estimates were computed.

Random aggregates of populations

Previous studies have examined case reports and incidence in both single populations and aggregate metapopulations. Here we construct random aggregates of various sizes. For each random aggregate, $M = X^2$ populations were sampled without replacement ($X \in 2, ..., 5$). Timeseries, total cases, and total susceptibles were summed over sampled cities. Reporting rate and incidence distribution was computed as above. For each M, 100 aggregates were drawn.

Estimating extinction boundary probabilities

To predict the distribution of ξ for a given population size, a separate linear model was fit to each metapopulation and inferred parameter using N as the independent variable (Table 2.2). We use these descriptive linear models to compute the per-week probability $B = Pr(\xi \leq \log \frac{1}{N})$ for a range of N.

The ECDF of ξ is clearly discrete, while the normal CDF is continuous. Nonetheless, we propose that B yields a good estimate of the amount of time each population spends at or below 1 actual case, providing an approximate *extinction boundary probability*. This estimate is based on inferred (or actual) rather than observed cases, and thus accounts for rescue effects and is not biased by apparent extinctions.

2.4 Results

Distribution of inferred incidence

Figure 2.1 shows weekly scaled case reports for a subset of U.S. cities, and the weekly unscaled mean of city case reports (omitting missing values). We use this dataset, along with the previously-studied England & Wales dataset and a simple stochastic metapopulation model, to show that the distribution of weekly per capita log incidence (ξ) provides a unifying framework for comparing populations and metapopulations.

Figure 2.2A (Data column) shows the inferred mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) for the per-city normal distribution of ξ. For comparison, the descriptive linear models of data are also plotted (see Table 2.2). Figure 2.2A (Model column) shows $\hat{\mu}$ and $\hat{\sigma}$ inferred from the best-fit epidemiological model ensembles. See Figure 2.13 for epidemiological model fit details.

The inferred $\hat{\sigma}$ of data (Figure 2.2A, left column) show more scatter in the U.S. than in England & Wales. One probable cause of this scatter is that reporting rate estimates appear less accurate in the U.S. (see Figure 2.12). Further, the geographical variation in the U.S. is extreme, where similarly-sized cities may range from close proximity to large cities to relative isolation. Lacking details of spatial connectivity, we explored simple measures of connectivity, including a rank gravity model, to explain the observed variance in inferred distributions. We did not identify a simple measure of connectivity that explains a significant proportion of this variance, though this area deserves more attention.

In simulation results (Figure 2.2A, right column), mean per capita incidence appears to saturate as population size increases. We might expect this because deterministic models of frequency-dependent transmission, where the force of infection

depends only on the fraction infected, predict that per capita incidence does not depend on population size. The data do not clearly exhibit the same saturation behavior, which suggests future modeling efforts should investigate more flexible assumptions about transmission.

Figure 2.2B shows the ratio of sample statistics (sample mean and standard deviation of observed log per capita incidence, excluding zeros) to their associated inferred parameters ($\hat{\mu}$ and $\hat{\sigma}$) for data. At small population sizes, sample statistics greatly underestimate the amount of variation observed in the data due to the exclusion of zero weeks ($\xi = -\infty$). Sample statistics converge towards inferred values at large population size. This population size, near prior estimates of CCS, is the threshold above which ξ is normally distributed.

ANCOVA results are shown in Table 2.1. Country identity has a significant effect on intercept (though not both slope and intercept). Overall, U.S. populations exhibit less persistence than comparably sized populations in England & Wales. Goodness-of-fit results are shown in Figures 2.9-2.11.

Revisiting Critical Community Size

For comparison with previous results [17, 20, 34], Figure 2.3A shows the relationship between population ($\log N$) and the relative frequency of zero weeks (P_0) for both data and models. Cities in England & Wales show lower P_0, which might suggest a lower CCS in cities in England & Wales than in the U.S. in this era. Yet these curves are not directly comparable because reporting rates and sampling frequency differ, affecting the probability of observing zeros (see Figure 2.7).

Figures 2.3B and C use model results to examine key processes that are not readily observable in real systems. For $\Gamma = E_w + I_w$, Figure 2.3B shows true extinctions ($\Gamma = 0$), apparent extinctions ($\Gamma > 0 \cap I_{wo} = 0$) and rescue events ($\Gamma = E_{w\eta}$) per

week. Rescue events and apparent extinctions are most common in intermediate-sized cities. These cities spend more time on the edge of extinction, where the influence of stochastic observational processes and imports are maximal. True extinctions, on the other hand, show clear curvilinear scaling with population size, following the trend evident in P_0 (Figure 2.3A). The number of observed zeros is equivalent to the sum of true and apparent extinctions. Thus, Figure 2.3B suggests that the scatter in observed zeros (Figure 2.3A), particularly in the U.S., is caused in part by apparent extinctions, while rescue events play a much smaller role.

Figure 2.3C displays apparent extinctions as a function of reporting rate, highlighting the interaction between reporting rate and population size. A clear constraint curve is evident, where the maximum apparent extinction rate is a function of the reporting rate. Increasing population size lowers the apparent extinction rate from this maximum towards zero for the largest populations.

Surprisingly, the effect of per-city reporting rates and the variability thereof has not been previously examined in detail. In a few cases, individual population estimates have been reported [15, 35], as well as overall metapopulation estimates [36, 37]. These generally agree with our findings of lower and more variable reporting rates in the U.S. than in England & Wales, though we found no previous estimates on within-metapopulation variance. Given the significant effect of variable reporting rates on observational bias shown in Figure 2.3C, this is a potentially fruitful avenue of study.

Figure 2.4 shows the effect of aggregation on the distribution of ξ. Random aggregates of smaller cities exhibit distributions of ξ similar to single, large cities, as shown by descriptive linear model predictions. In the US, aggregate $\hat{\mu}$ is consistently above the linear model prediction, while $\hat{\sigma}$ is consistently below the prediction. Thus aggregation consistently reduces variation in the US, which would be expected amongst asynchronous populations. In England & Wales, the pattern is less clear-cut, with

aggregates falling both above and below the linear model prediction for both inferred parameters. The above is consistent with the observation that the mean pairwise population correlation coefficient is much higher in England & Wales (0.29) than in the U.S. (0.15), indicating greater asynchrony in the U.S.

The observation that random aggregates generally follow the patterns of ξ predicted by linear models of single populations argues for a reconsideration of single large cities as key drivers in disease persistence in general and the emergence of measles in particular. Previous work used CCS to infer the possible historical era of measles zoonosis based on historical population sizes [19]. Our results argue strongly in favor of a metapopulation-level view of disease emergence, where interconnected aggregates of small populations can support disease persistence. This fact has important modern implications for zoonosis, which often occurs at the interface between human settlements and natural systems.

As the size of random aggregates grows, a central limiting distribution of ξ is reached, such that per capita log incidence is constant with increasing population size. This is the expected behavior for frequency-dependent transmission and is suggested by epidemiological model results, as mentioned above. This limiting distribution is very different for the U.S. and England & Wales, with the U.S. exhibiting relatively smaller $\hat{\mu}$ and larger $\hat{\sigma}$.

The *extinction boundary probability* $B = Pr(\xi \leq \log \frac{1}{N})$, as computed from $\hat{\mu}$ and $\hat{\sigma}$, is plotted for each population and random aggregate in Figure 2.4B and C. Intuitively, B forms an upper boundary for the per-week probability of a population being in the extinct state. The plotted curves show B for a range of population sizes N, as estimated from the descriptive linear models predicting $\hat{\mu}$ and $\hat{\sigma}$ from individual city size (fit for each metapopulation, see Table 2.2). Thus, for $N > 10^7$, we expect fewer than one extinction in a thousand years ($B < 10^{-5}$) in both metapopulations. Figures 2.4B and C are identical except for log scaling of the Y axis in C.

Figures 2.4B and C highlight the difference between random aggregate and metapopulation estimates. In the U.S., random aggregates more closely match metapopulation estimates than in England & Wales. In addition, U.S. random aggregates are generally below (e.g. less likely to be extinct) what would be predicted from the metapopulation curve, as would be expected from the aggregation of asynchronous populations.

The extinction boundary probabilities give rise to a probabilistic interpretation of CCS. A given probability B_α has a corresponding critical community size CCS_α; for populations larger than CCS_α, the predicted per-week probability of being extinct is less than α. Indeed, the metapopulation curves reveal that B is higher in the U.S. than in England & Wales for all population sizes, yielding a larger CCS_α for all α.

Epidemiological model results

Even with a wealth of case report and demographic data, we still lack sampling of rural populations and patterns of spatial connectivity. We have thus chosen a relatively parsimonious epidemiological model formulation here that implicitly includes space. Despite the simple formulation, simulations do capture the overall scaling of $\hat{\mu}$ and $\hat{\sigma}$, as well as the proportion of observed zeros P_0 (Figure 2.3A). Representative simulation timeseries are shown in Figure 2.14. Nonetheless, the observed scaling of $\hat{\mu}$ and $\hat{\sigma}$ with population size N (Figure 2.2A) is not fully reproduced. Simulations yield nonlinear scaling, with per capita incidence approximately constant above a threshold N. Note, however, that Figure 2.2A shows ensemble means, which greatly reduces between-population scatter compared to individual simulations. We suggest that inferred $\hat{\mu}$ and $\hat{\sigma}$ provide important probes that more complex mechanistic models can be tested against.

Epidemiological model results also clearly illustrate the influence of key dynamical

processes that are difficult to observe in real systems (Figure 2.3B). Here we find that apparent extinctions greatly affect observational processes in midsized cities, an effect that is compounded by the low reporting rates of the US. A range of parameter values were found to produce similar results (see Figure 2.13). The final epidemiological models presented here are primarily for illustrative purposes, and different parameter choices do not affect overall results.

2.5 Discussion

Comparing metapopulations

The U.S. dataset presented here provides an important counterpoint to the highly successful England & Wales measles dataset [13, 14, 16, 38, 39]. First and foremost, it contains extensive spatial and temporal heterogeneity compared to England & Wales. Demographics and transportation vary over time, from boom years (1920s) to economic depression (1930s) to a major war and demographic boom (1940s), accompanied by racial and ethnic segregation within and between cities [28, 40, 41]. Population density and transportation networks vary greatly in space, from dense and highly-connected Northeastern cities to isolated mountain West communities such as Billings, Montana and the island community of Galveston, Texas.

Both datasets sample a single, extensive metapopulation over a long, contiguous period of time. England & Wales offers dense coverage of a spatially compact metapopulation, while the U.S. dataset samples a much larger metapopulation, albeit less densely. The U.S. in this era is socially, economically, and even genetically more heterogeneous than England & Wales. In addition, key drivers of measles dynamics, such as school terms and family size, vary greatly throughout the U.S. during this era [42], both temporally and spatially. Lacking data on these factors, we have used a

simple "strategic" rather than detailed "tactical" model formulation that nonetheless largely reproduces observed patterns in the data.

Here we argue that the inferred distribution of weekly per capita log incidence ξ, and its consistent scaling with population size N, yields a robust comparison of patterns of incidence between metapopulations. Because our method accounts for reporting rate, population size and sampling frequency, we suggest that the remaining differences between metapopulation distributions result from differences in underlying ecological processes. The results presented here show significant differences in the distribution of ξ between countries (Table 2.1), as well as in the limiting distribution of ξ in random aggregates (Figure 2.4). Several key points deserve special notice. First, reporting rates are lower and more variable in the U.S. (Figure 2.3C), which suggests a larger apparent CCS, and which we have corrected for. Second, higher birth rates in the U.S. would generally favor increased persistence and a lower CCS, which we do not observe. Finally, large intercity distances in the U.S. could result in a smaller rescue effect, leading to a larger true CCS, as observed here.

Distributional measure of persistence and extinction

Diseases with relatively high R_0 and short infectious period, such as measles, are prone to local fade-outs. Large focal communities have been proposed as refugia that allow metapopulation persistence, thus highlighting the importance of CCS. Yet the usefulness of fixed population thresholds, such as CCS, has been criticized: Lloyd-Smith et al. [43] points out that "thresholds are rarely abrupt and always difficult to measure". By taking a distributional approach to understanding persistence and extinction in host-pathogen interactions, we aim to bypass some of the shortcomings of single threshold measures.

Our choice of log incidence (ξ) is motivated by the observation that incidence of

acute infectious disease generally emerges from a multiplicative process, where the natural scale is logarithmic [44]. However, a log scale does not easily permit consideration of zero incidence and, in these situations, the true distribution of ξ is often thought of as a mixture of one process that describes presence/absence and another that conditions on presence and describes non-zero incidence [45]. Yet observations of zero cases do not imply zero incidence in this system. Thus we subsume observed zeros into incidence at or below the minimum observable nonzero incidence, $\xi = \log \frac{1}{rN}$. This approach preserves the within-city empirical CDF (ECDF) by including all the observed data points, yet is not overly biased by our inability to observe incidence between 0 and $\frac{1}{rN}$ ($\xi \in (-\infty, \log \frac{1}{rN})$).

Theory based on simple non-seasonal stochastic epidemiological models predicts that, for large enough population size, the distribution of infectives is approximately normal [46, 47], with mean and variance scaling linearly with population size. These results are obtained by conditioning on non-extinction to derive a quasi-stationary distribution of infectives. In fact, for average measles parameters, the threshold for this approximation can be close to 10^7, which is above the size of all cities in the two metapopulations considered here. For both datasets, we find that the normal CDF is a good description of the marginal distribution of log incidence for populations at or over the critical community size, so that sample mean and standard deviation can be used to infer this distribution (as shown in Figure 2.2). We argue that this is empirical evidence that the time series can be considered as weakly stationary, despite the intrinsic seasonality and autocorrelation present in the system. This supports theoretical work demonstrating that, if populations are above a critical size, the assumption that infectives follow a lognormal distribution is valid when closing moment equations [48, 49]. In addition, for smaller populations, we can still characterize log incidence using a normal distribution by fitting a normal CDF to the truncated empirical CDF. Understanding how this inferred distribution relates to recent work by Black and McKane [50], who derived an analytic approximation to

the marginal distribution of infectives for a non-seasonal SIR model, could provide further insight into why the lognormal is a good fit for a wide range of population sizes.

Generality of approach and broader applications

The sensitivity of inferred distributions to the duration of study remains an important question. In particular, how long must populations be observed before log incidence is well-estimated by a normal distribution, and how do the inferred distributions change over time? The results presented here broadly hold if the data are divided into a small number of equal-length subdivisions (for details, see Figure 2.12). As subdivisions grow shorter, however, estimates diverge from those based on the full timeseries, and variance between subdivisions exceeds variance between populations within a subdivision. This is likely due to the fact that we are not able to observe enough of the distribution in a single short time series.

The above analysis assumes that processes in the studied metapopulations are weakly stationary over the period of examination. As such, the period of analysis should not include large perturbations, such as the onset of vaccination. This points to one potential application of this method: comparing metapopulation dynamics within a population pre- and post-vaccination. Thus metapopulation spatial structure remains static while effective birth rates decrease. In this case, both birth rates and vaccine uptake rates are required to estimate per-population reporting rate, and increased error is likely introduced. Nonetheless, this method could provide a direct measure of the efficacy of vaccination on incidence reduction.

Another outstanding question is whether this approach extends from measurements of incidence (e.g. per capita new cases in a unit of time) to more ecologically common measures of prevalence (e.g. per capita cases or occurrences at an instant

in time). However, incidence and prevalence are very similar for diseases that are reported at intervals close to the average infectious period. For example, if we assume that incidence represents all newly recovered individuals over a reporting interval s then incidence $\xi = \gamma \int_t^{t+s} I(t)dt$, where $1/\gamma$ is the average infectious period. If $s = 1/\gamma$ then incidence is the average prevalence over that interval.

Conversion between incidence and prevalence is important for several reasons. Prevalence is the ecologically relevant measure of disease burden, even though incidence is typically measured in human diseases (as case reports). On the other hand, in many ecological systems including non-human diseases, prevalence is the only obtainable measurement. We suggest the method presented is applicable to measurements of either prevalence or incidence, provided the sampling interval is short compared to the generation time and dynamics of the studied disease and host population.

Unraveling population and metapopulation dynamics from chance and observational processes is a difficult question in many systems. Stochasticity, spatial connectivity, and incomplete observation each play important and interconnected roles in this system. Here we present a framework that concisely infers the marginal distribution of measles incidence within populations. Comparison of inferred distributions between populations yields a high-level picture of metapopulation incidence patterns. The result is a probabilistic measure of persistence that can be used to compare and unify ecological models, data, and theory.

Whether the results shown here generalize to other systems remains an open question. The explosive population dynamics of measles are distinct but not unique. Influenza, pertussis, and polio, for example, exhibit epidemic peaks as well as immunity-mediated demographic extinction (though influenza's rapid evolution precludes an estimation of reporting rate by the approach used here). More broadly, wildlife management is one field where metapopulation theory has long been applied, and

where estimates of incidence distributions may prove useful. The establishment of an invasive species and the preservation of a threatened population are analogous to pathogen emergence and persistence. For infectious diseases, the host population is analogous to patch area, and prevalence becomes analogous to population density. Inferred distributions of incidence (or prevalence) permit the direct estimation of extinction or persistence risk from existing time series of population sizes or densities, and provide simple measures that link population and metapopulation ecology.

Acknowledgments

The authors would like to thank Stacy O. Scholle, Etsuko Nonaka, Duncan Wadsworth, John Hammond, Erik Erhardt, and Natalie Wright for suggestions, ideas, and encouragement, and the Wearing Lab data entry team for their hard work. Joe Conway assisted with database design. The editors and anonymous reviewers provided thoughtful comments that contributed greatly to the manuscript.

CG was supported by a fellowship in the Program in Interdisciplinary Biological and Biomedical Sciences at the University of New Mexico. This publication was made possible by Grant Numbers P20RR018754 from the National Center for Research Resources (NCRR), T32EB009414 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and U01GM09766 from the National Institute of General Medical Sciences (NIGMS), components of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR, NIBIB, NIGMS or NIH.

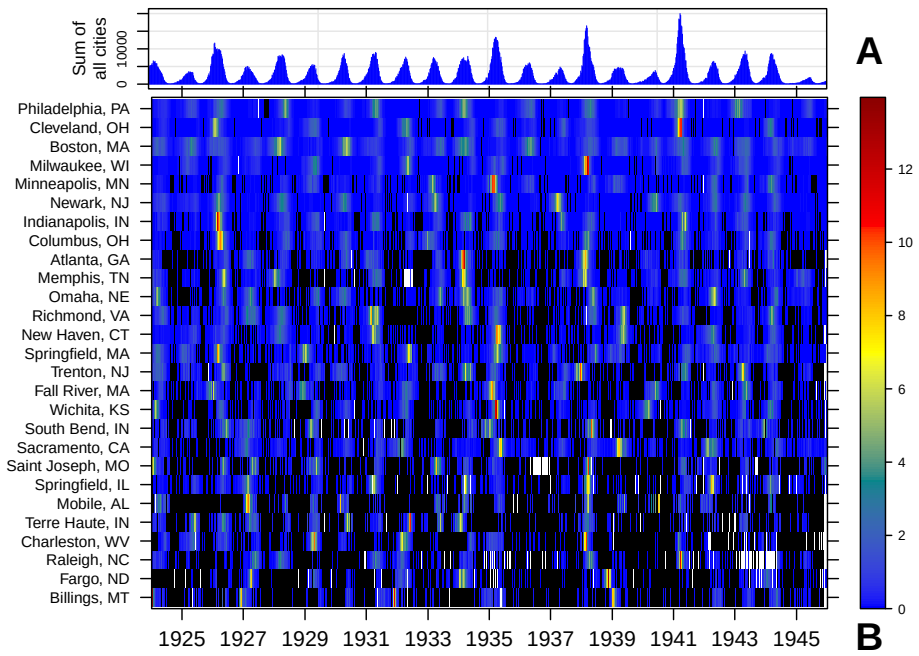

Figure 2.1. Weekly measles case reports for a subsample of United States cities, 1924-01-05 to 1945-12-29. Zeros are black and missing values are white. **(A)** Weekly sum of unscaled case reports of all cities. **(B)** Heatmap for a subset of cities ordered by mean population size. Values are variance-scaled within each city.

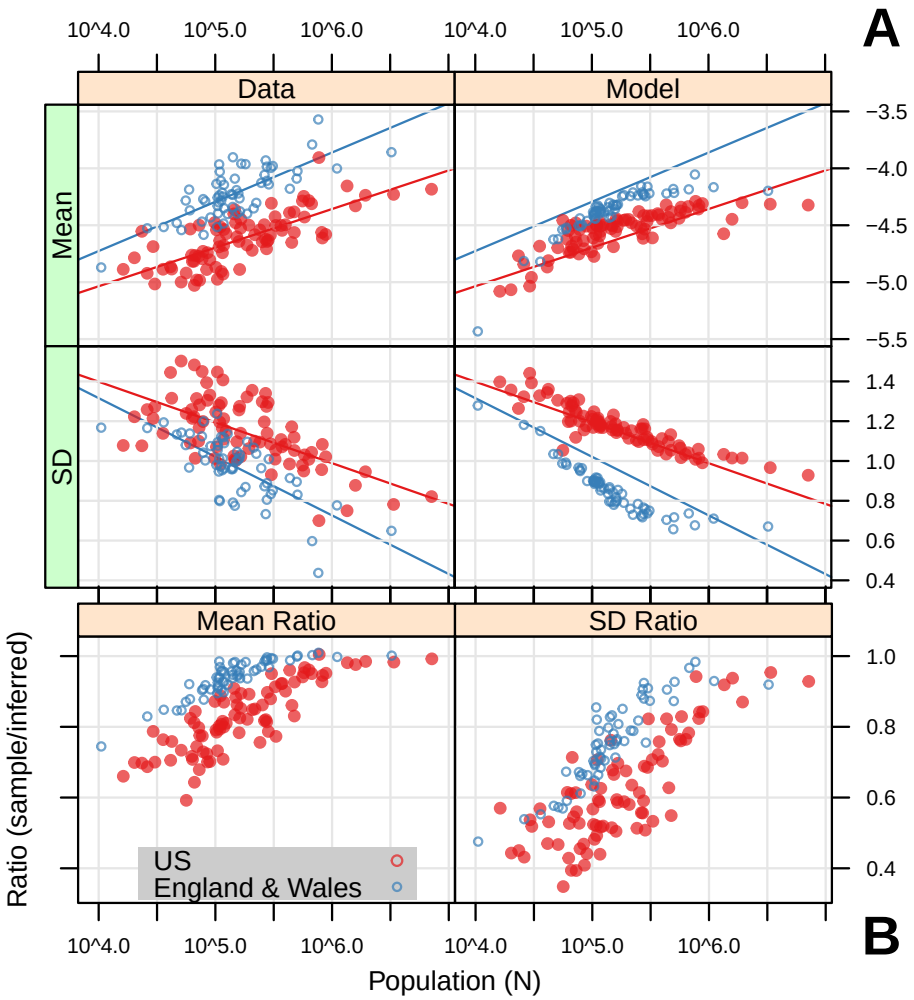

Figure 2.2. Inferred parameters for normal distributions of weekly log incidence (ξ). (A) Left column: for each city, a normal CDF with mean ($\hat{\mu}$) and SD ($\hat{\sigma}$) was fit to the ECDF of ξ by nonlinear minimization using an L_∞ metric (minimizing Kolmogorov-Smirnoff (K-S) distance). Right column: Simulation results, with ($\hat{\mu}$) and SD ($\hat{\sigma}$) inferred as above, averaged over 50 realizations. See Table 2.5 for final epidemiological model parameters. For comparison, descriptive linear models of inferred parameters against $\log N$ of data (left column) are shown in both columns (see Table 2.2). **(B)** Ratio of sample statistics to inferred parameters (sample/inferred) for data. Sample statistics underestimate variation at small population sizes, and converge towards inferred parameters at large population sizes.

Inferred Parameter	Model Term	Estimate	Std. Error	t value	Pr(¿—t—)
Mean	(Intercept)	-6.327	0.148	-42.842	$< 10^{-12}$
Mean	logN	0.367	0.028	12.967	$< 10^{-12}$
Mean	Country	-0.210	0.013	-16.021	$< 10^{-12}$
SD	(Intercept)	2.261	0.118	19.222	$< 10^{-12}$
SD	logN	-0.231	0.023	-10.263	$< 10^{-12}$
SD	Country	0.095	0.010	9.057	$< 10^{-12}$

Table 2.1. ANCOVA results. A separate linear model was constructed for each inferred parameter, using log N and country identity as predictors. Simulations were not modelled. England & Wales is the reference level. Mean ($\hat{\mu}$), $R^2 = 0.74$; SD ($\hat{\sigma}$), $R^2 = 0.55$. For both inferred parameters, country identity has a significant impact on either intercept (shown here) or slope, but not both. Note the model is not balanced.

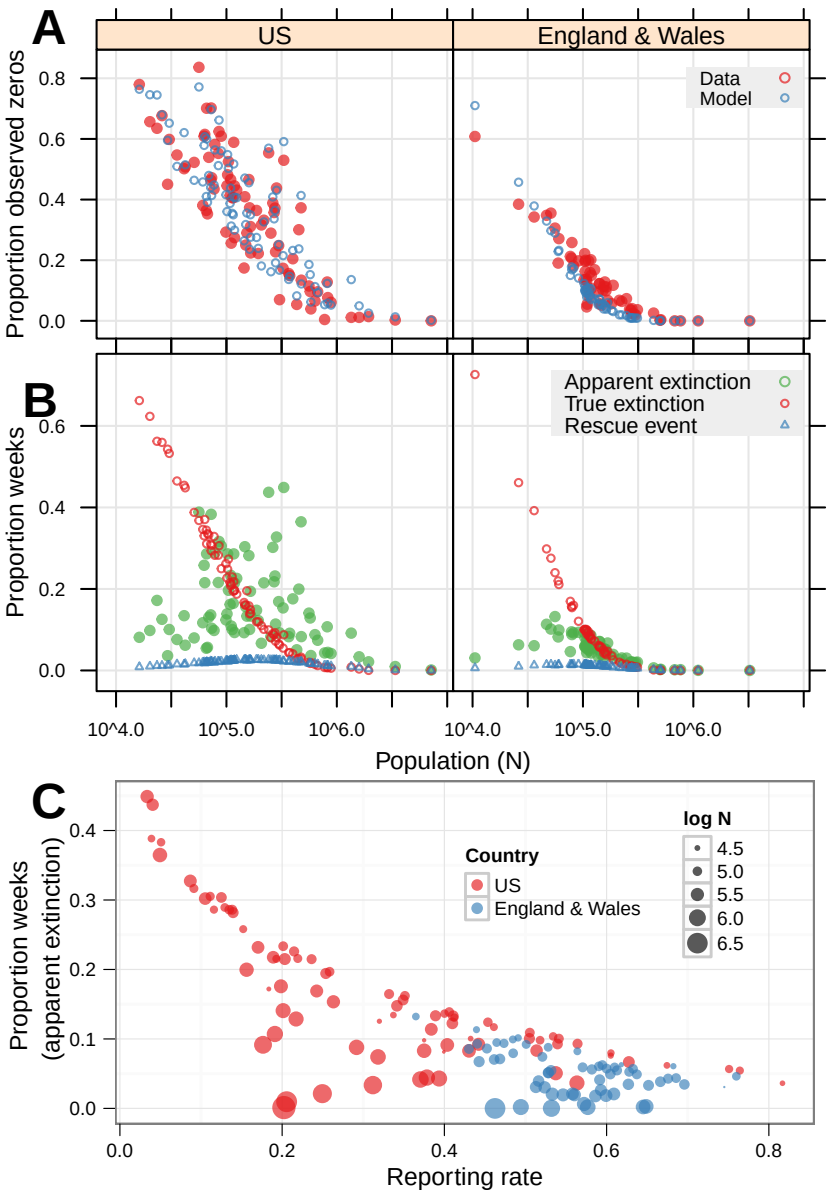

Figure 2.3. Distribution of zeros, extinctions, and rescues. For epidemiological models, values were calculated for each realization and an ensemble mean taken. **(A)** Proportion of observed zeros by population size. **(B)** Proportion of apparent extinctions ($\Gamma > 0 \cap I_{wo} = 0$), true extinctions ($\Gamma = 0$), and rescue events ($\Gamma = E_{w\eta} > 0$) for $\Gamma = E_w + I_w$. True extinctions scale with population size; rescue events and apparent extinctions are highest at intermediate populations. Apparent extinctions cause scatter in observed zeros (A) in the U.S. **(C)** Apparent extinction versus reporting rate. As population size decreases, the apparent extinction rate becomes more sensitive to reporting rate.

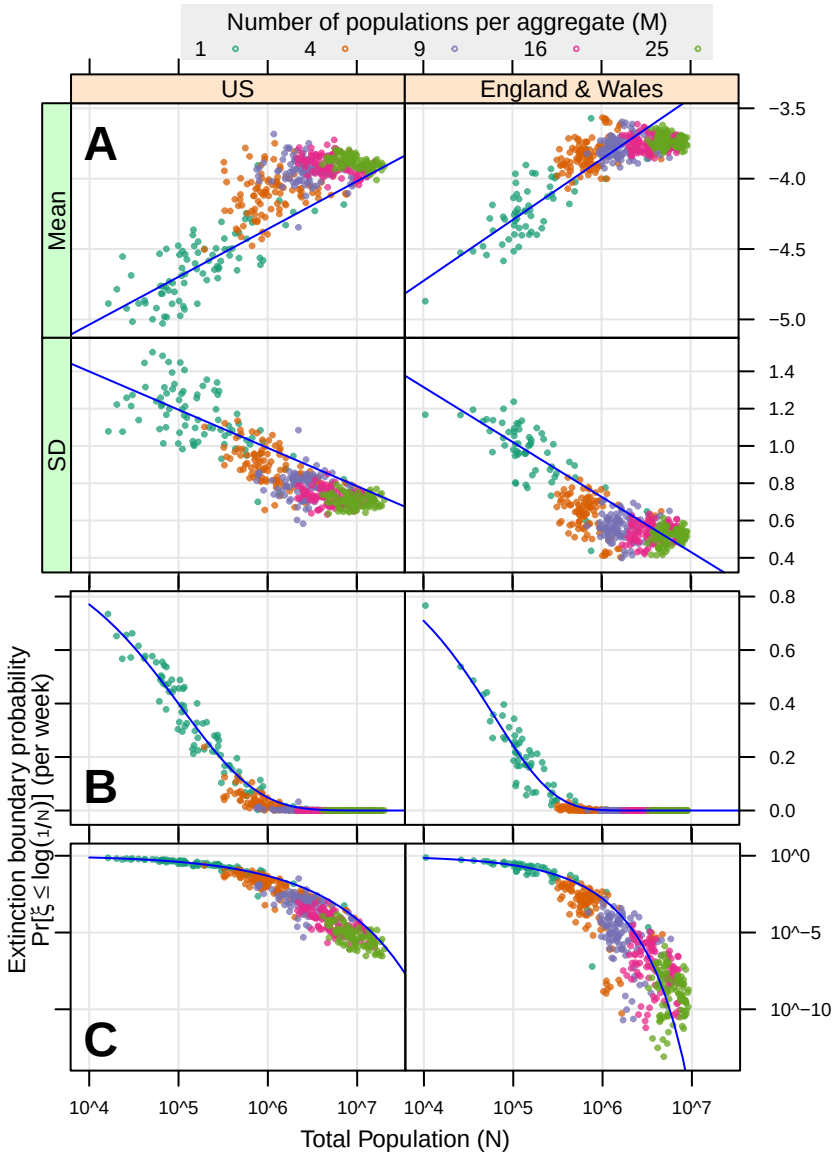

Figure 2.4. Inferred distributions (A) and extinction boundary probabilities (B,C) for single populations and random aggregates of populations. 100 random aggregates were drawn for each aggregate size (see legend). **(A)** Inferred distributions and linear models (linear models exclude random aggregates, see Table 2.2). For random aggregate total population $N \sim\geq 10^7$, ξ converges to a limiting distribution. **(B, C)** Points show the *extinction boundary probability* $B = Pr(\xi \leq \log \frac{1}{N})$, estimated from $\hat{\mu}$ and $\hat{\sigma}$ for populations and random aggregates in (A). Curves show B for a range of population sizes, as predicted by linear models from (A). Any probability $B = \alpha$ has a corresponding population size, giving a probabilistic measure of critical community size CCS_α. The U.S. B curve is higher than in England & Wales, indicating higher probabilities of extinction across population sizes.

2.6 References

[1] A.A. King, S. Shrestha, E.T. Harvill, and O.N. Bjørnstad. Evolution of Acute Infections and the Invasion-Persistence Trade-Off. *Am. Nat.*, 173(4):446–455, 2009.

[2] W.H. Hamer. Epidemic disease in England. *Lancet*, 1, 1906.

[3] H.E. Soper. The interpretation of periodicity in disease prevalence. *J R Stat Soc*, 92(1):34–73, 1929.

[4] M.S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 4, pages 81–109. University of California Press Berkeley, 1956.

[5] F.L. Black. Measles endemicity in insular populations: critical community size and its evolutionary implication. *J. Theor. Biol.*, 11(2):207–211, 1966.

[6] R.M. Anderson and R.M. May. *Infectious diseases of humans*. Oxford University Press Oxford, 1991.

[7] M.J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[8] M.J. Ferrari, R.F. Grais, N. Bharti, A.J.K. Conlan, O.N. Bjørnstad, L.J. Wolfson, P.J. Guerin, A. Djibo, and B.T. Grenfell. The dynamics of measles in sub-Saharan Africa. *Nature*, 451(7179):679–684, 2008.

[9] B.M. Bolker and B.T. Grenfell. Chaos and biological complexity in measles dynamics. *Proceedings: Biological Sciences*, pages 75–81, 1993.

[10] D.J.D. Earn, P. Rohani, B.M. Bolker, and B.T. Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453):667, 2000.

[11] M.J. Keeling, P. Rohani, and B.T. Grenfell. Seasonally forced disease dynamics explored as switching between attractors. *Physica D*, 148(3-4):317–335, 2001.

[12] P. Rohani, M.J. Keeling, and B.T. Grenfell. The interplay between determinism and stochasticity in childhood diseases. *Am. Nat.*, 159(5):469–481, 2002.

[13] B. Bolker and B. Grenfell. Space, persistence and dynamics of measles epidemics. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 348(1325):309–320, 1995.

[14] B.T. Grenfell and B.M. Bolker. Cities and villages: infection hierarchies in a measles metapopulation. *Ecol. Lett.*, 1(1):63–70, 1998.

[15] W.P. London and J.A. Yorke. Recurrent outbreaks of measles, chickenpox and mumps: I. Seasonal variation in contact rates. *Am. J. Epidemiol.*, 98(6):453, 1973.

[16] P.E.M. Fine and J.A. Clarkson. Measles in England and Wales. I. An analysis of factors underlying seasonal patterns. *Int J Epidemiol*, 11(1):5–14, 1982.

[17] M.S. Bartlett. Measles periodicity and community size. *J R Stat Soc Ser A*, 120(1):48–70, 1957.

[18] I. Nåsell. A new look at the critical community size for childhood infections. *Theor Popul Biol*, 67(3):203–216, 2005.

[19] A.J.K. Conlan and B.T. Grenfell. Seasonality and the persistence and invasion of measles. *Proc. R. Soc. B*, 274(1614):1133–1141, 2007.

[20] A.J.K. Conlan, P. Rohani, A.L. Lloyd, M. Keeling, and B.T. Grenfell. Resolving the impact of waiting time distributions on the persistence of measles. *J R Soc Interface*, 7(45):623, 2010.

[21] B.T. Grenfell, O.N. Bjørnstad, and B.F. Finkenstädt. Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. *Ecol Monogr*, 72(2):185–202, 2002.

[22] D.A. Griffiths. The effect of measles vaccination on the incidence of measles in the community. *J R Stat Soc Ser A*, pages 441–449, 1973.

[23] B.M. Bolker and B.T. Grenfell. Impact of vaccination on the spatial correlation and persistence of measles dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 93(22): 12648–12653, 1996.

[24] W.A. Orenstein, K.L. Samuel, and A.R. Hinman. Summary and conclusions: measles elimination meeting, 16–17 march 2000. *J. Infect. Dis.*, 189(Suppl 1): S43–S47, 2004.

[25] I. Hanski, A. Moilanen, and M. Gyllenberg. Minimum viable metapopulation size. *Am. Nat.*, pages 527–541, 1996.

[26] U.S. Public Health Service. Public Health Rep, 1920–1950. URL `http://www.ncbi.nlm.nih.gov/pmc/issues/149156/`.

[27] U.S. Census Bureau. U.S. Census, 1920–1950.

[28] U.S. Census Bureau. Statistical Abstract of the United States, 1920–1950.

[29] P. Rohani. Personal communication, 2012.

[30] H. Southall. A vision of Britain through time: making sense of 200 years of census reports. *Local Popul Stud*, 76:76, 2006.

[31] J.E. Dennis and R.B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice-Hall, 1983.

[32] L.F. Olsen, G.L. Truty, and W.M. Schaffer. Oscillations and chaos in epidemics: a nonlinear dynamic study of six childhood diseases in Copenhagen, Denmark. *Theor Popul Biol*, 33(3):344–370, 1988.

[33] D.T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys*, 115:1716, 2001.

[34] M.S. Bartlett. The critical community size for measles in the United States. *J R Stat Soc Ser A*, 123(1):37–44, 1960.

[35] J.A. Clarkson and P.E.M. Fine. The efficiency of measles and pertussis notification in England and Wales. *Int J Epidemiol*, 14(1):153–168, 1985.

[36] F.L. Black. The role of herd immunity in control of measles. *Yale J Biol Med*, 55(3-4):351, 1982.

[37] B.F. Finkenstädt and B.T. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *J R Stat Soc Ser C Appl Stat*, 49(2):187–205, 2000.

[38] P. Rohani, D.J.D. Earn, and B.T. Grenfell. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286(5441):968, 1999.

[39] B.T. Grenfell, O.N. Bjørnstad, and J. Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.

[40] S.E. Tolnay. The African American Great Migration and Beyond. *Annu Rev Sociol*, pages 209–232, 2003.

[41] W.D. Middleton, G.M. Smerk, and R.L. Diehl. *Encyclopedia of North American Railroads*. Indiana Univ Pr, 2007.

[42] B. Metzker. School Calendars. Technical Report EDO-EA-02-03, Educational Resources Information Center, 2002.

[43] J.O. Lloyd-Smith, P.C. Cross, C.J. Briggs, M. Daugherty, W.M. Getz, J. Latto, M.S. Sanchez, A.B. Smith, and A. Swei. Should we expect population thresholds for wildlife disease? *Trends Ecol. Evol.*, 20(9):511–519, 2005.

[44] E. Limpert, W.A. Stahel, and M. Abbt. Log-normal distributions across the sciences: keys and clues. *BioScience*, 51(5):341–352, 2001.

[45] D. Fletcher, D. MacKenzie, and E. Villouta. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environ Ecol Stat*, 12(1):45–54, 2005.

[46] I. Nåsell. On the time to extinction in recurrent epidemics. *Proc. R. Soc. B*, 61 (2):309–330, 1999.

[47] H. Andersson and T. Britton. Stochastic epidemics in dynamic populations: quasi-stationarity and extinction. *J Math Biol*, 41(6):559–580, 2000.

[48] M.J. Keeling. Multiplicative moments and measures of persistence in ecology. *J. Theor. Biol.*, 205(2):269–281, 2000.

[49] A.L. Lloyd. Estimating variability in models for recurrent epidemics: assessing the use of moment closure techniques. *Theor Popul Biol*, 65(1):49–65, 2004.

[50] A.J. Black and A.J. McKane. WKB calculation of an epidemic outbreak distribution. *J Stat Mech*, 2011:P12006, 2011.

2.7 Appendix S1 – Additional Figures

Country	Inferred Parameter	Model Term	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
US	Mean	(Intercept)	-6.392	0.179	-35.710	$< 10^{-9}$
US	Mean	logN	0.339	0.034	9.963	$< 10^{-9}$
US	SD	(Intercept)	2.217	0.151	14.712	$< 10^{-9}$
US	SD	logN	-0.205	0.029	-7.143	$< 10^{-9}$
England & Wales	Mean	(Intercept)	-6.457	0.263	-24.579	$< 10^{-9}$
England & Wales	Mean	logN	0.432	0.051	8.532	$< 10^{-9}$
England & Wales	SD	(Intercept)	2.494	0.188	13.276	$< 10^{-9}$
England & Wales	SD	logN	-0.295	0.036	-8.125	$< 10^{-9}$

Table 2.2. Descriptive linear models (separated in table by horizontal lines) for each country and inferred parameter against log N. This gives a closed-form expression for the CDF of log incidence as a function of N, which was used to compute the *extinction boundary probability* B for a range of population sizes.

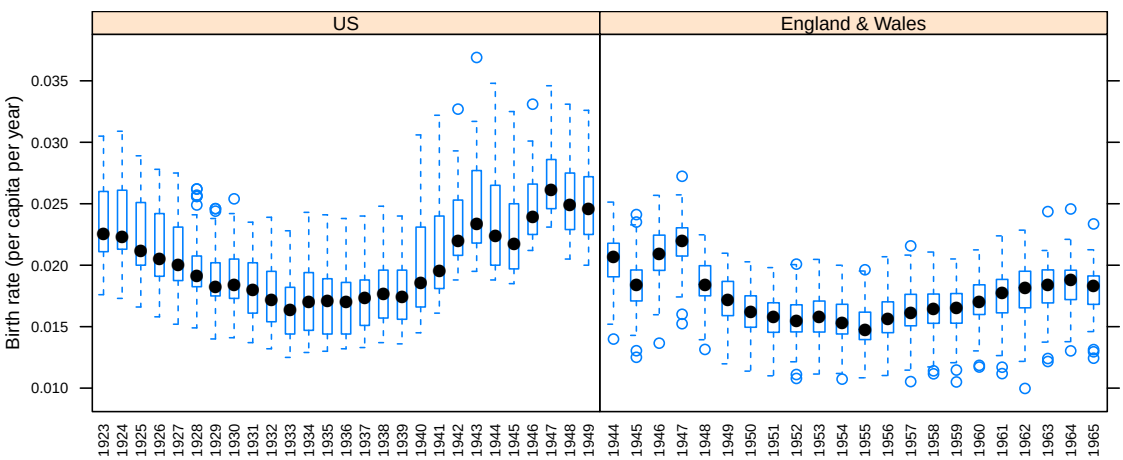

Figure 2.5. Birth rates adjusted for infant mortality. All rates are yearly per capita. The U.S. figure shows state birth rates, while the England & Wales figure shows city rates.

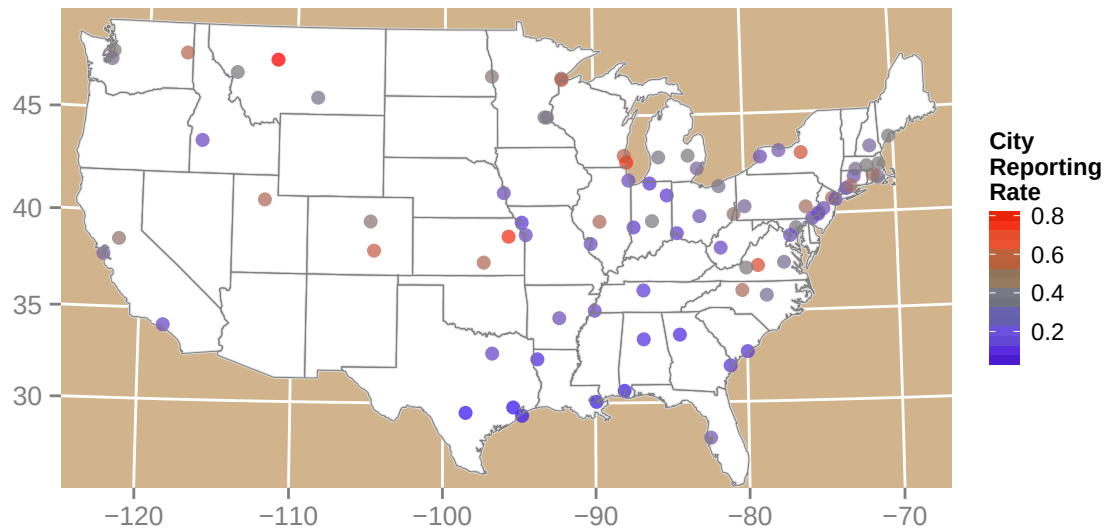

Figure 2.6. Sampled U.S. cities, showing estimated reporting rates.

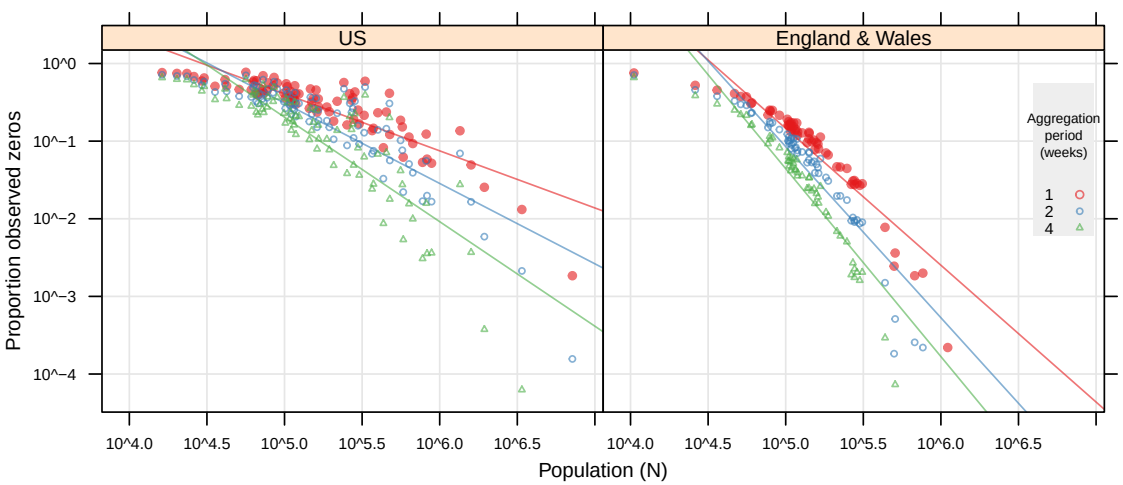

Figure 2.7. Classic CCS curve for varying levels of temporal aggregation. Observed weekly cases from simulations were summed over varying number of weeks. The mean proportion of observed zeros per sampling period was computed for each city. Linear regressions for each sample period are overlaid.

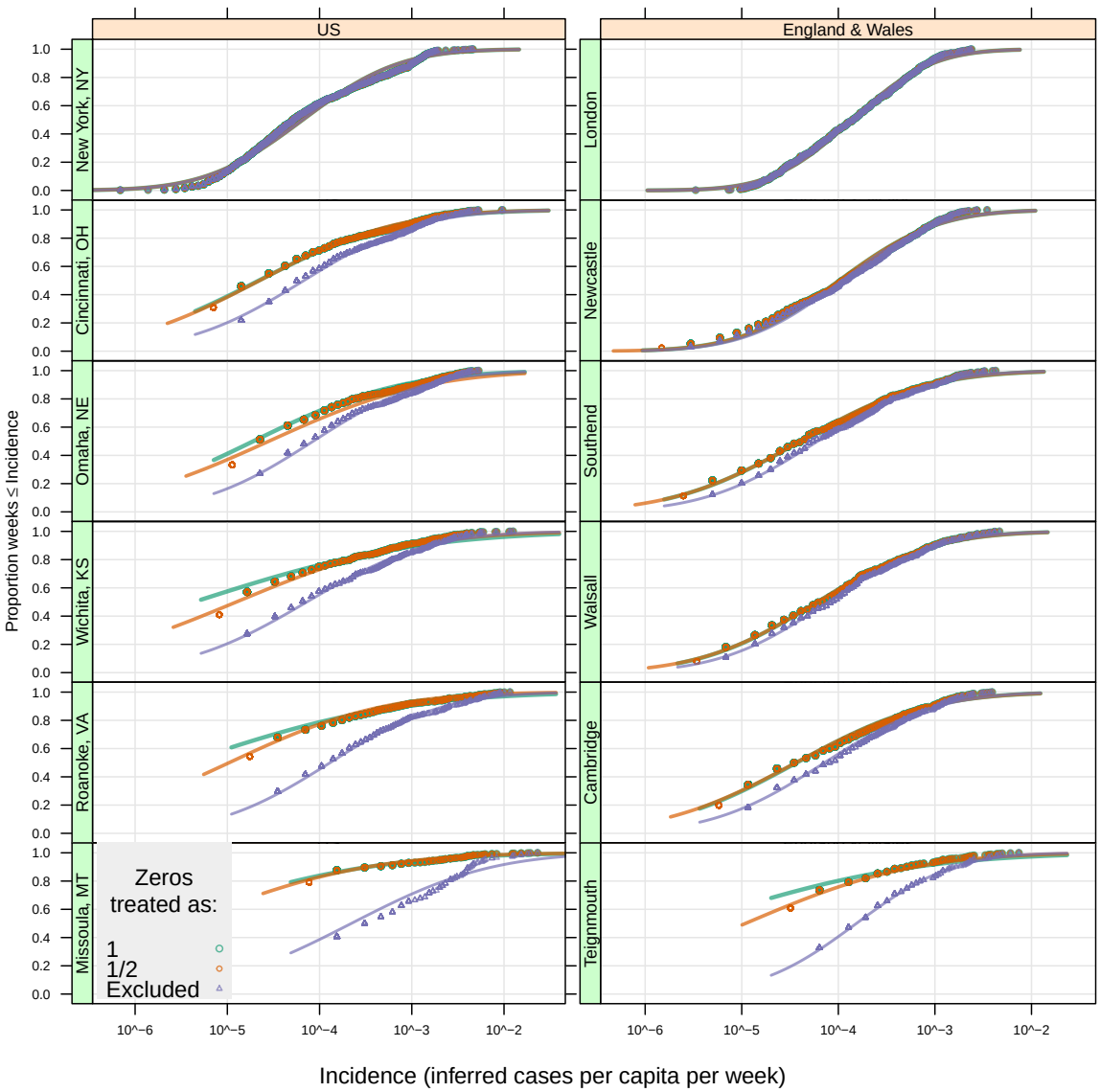

Figure 2.8. The ECDF (points) and fitted normal distribution (lines) of log incidence (ξ) for a sample of (population-ordered) cities. Here zeros are treated either as 1 or 1/2 actual case, or are excluded. Excluding zeros significantly shifts the ECDF and resulting fit, particularly in the U.S. Discretization of the ECDF is clearly evident at low incidence.

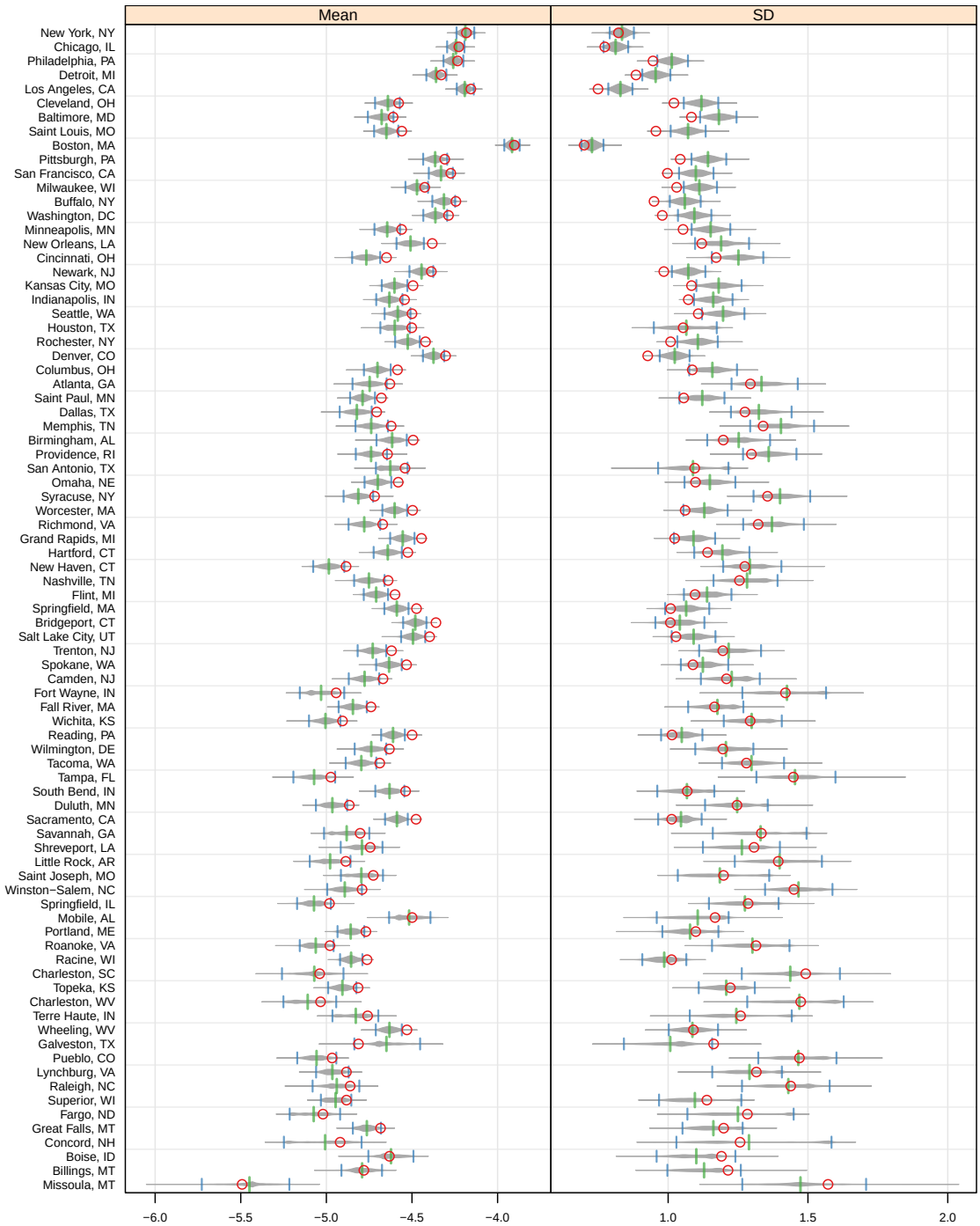

Figure 2.9. U.S. distribution of parameters inferred from parametric bootstrapped case reports. Red circle shows value inferred from data. Central green line shows bootstrap mean. Blue lines show 95% confidence interval. See Figure 2.11 for details.

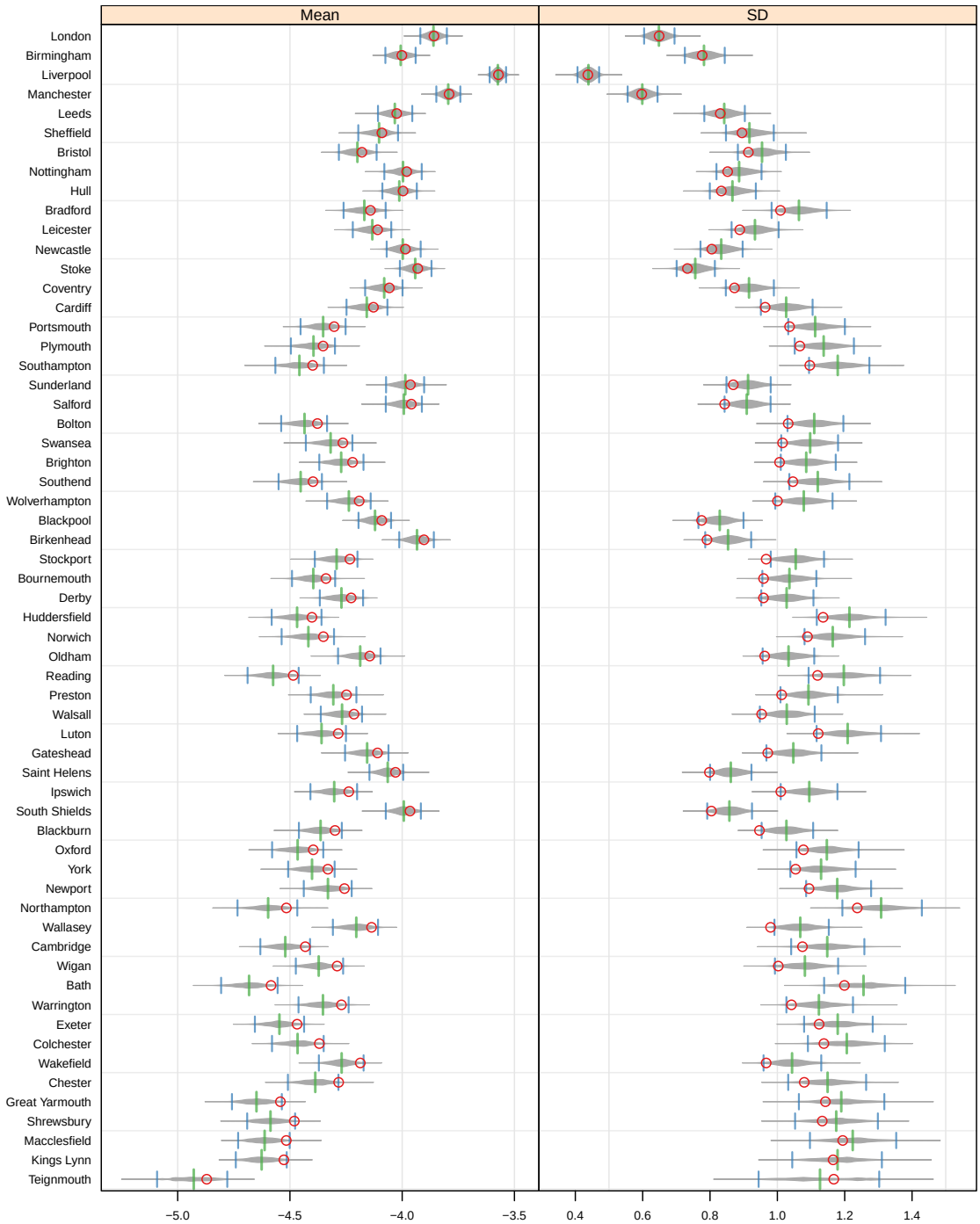

Figure 2.10. England & Wales distribution of parameters inferred from parametric bootstrapped case reports. Red circle shows value inferred from data. Central green line shows bootstrap mean. Blue lines show 95% confidence interval. See Figure 2.11 for details.

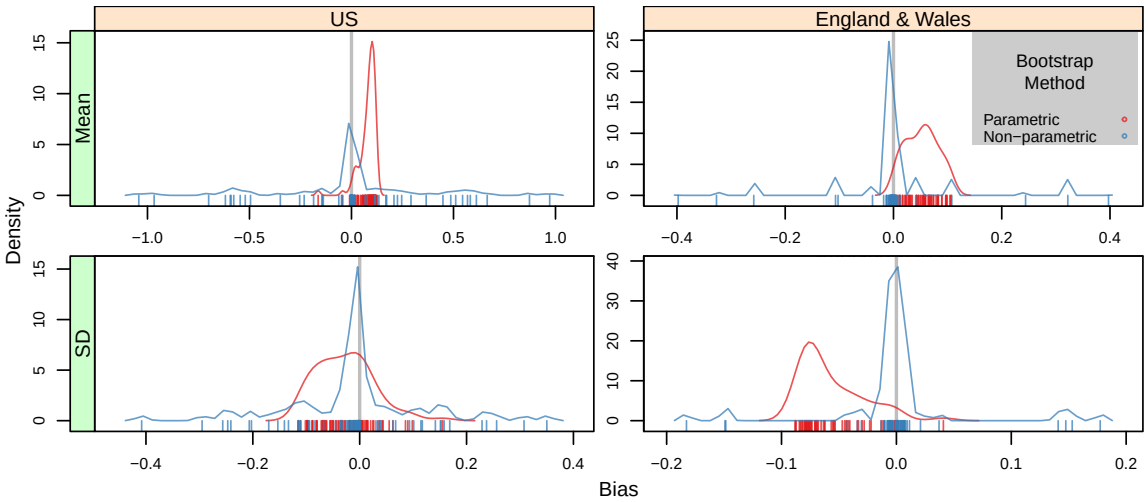

Figure 2.11. Bias of inferred parameters, as estimated from synthetic (bootstrap) datasets of case reports. For non-parametric bootstraps, case reports were sampled with replacement from original timeseries, excluding NAs. For parametric bootstraps, inferred cases were constructed by sampling ξ_i from $N(\hat{\mu}_i, \hat{\sigma}_i)$. Each ξ_i was anti-logged and multiplied by N_i to yield $\hat{C}_i$. C_i were obtained from $\hat{C}_i$ via binomial sampling, with a per-trial success rate of r_i. One thousand replicates were evaluated for each process. Bias is the distance between the bootstrap mean and the value inferred from data. Parametric bootstraps show consistent bias with lower variation of bias. The mean bias of non-parametric bootstraps is approximately zero, with very high bias for a small number of populations. See Figures 2.9 and 2.10 for parametric results.

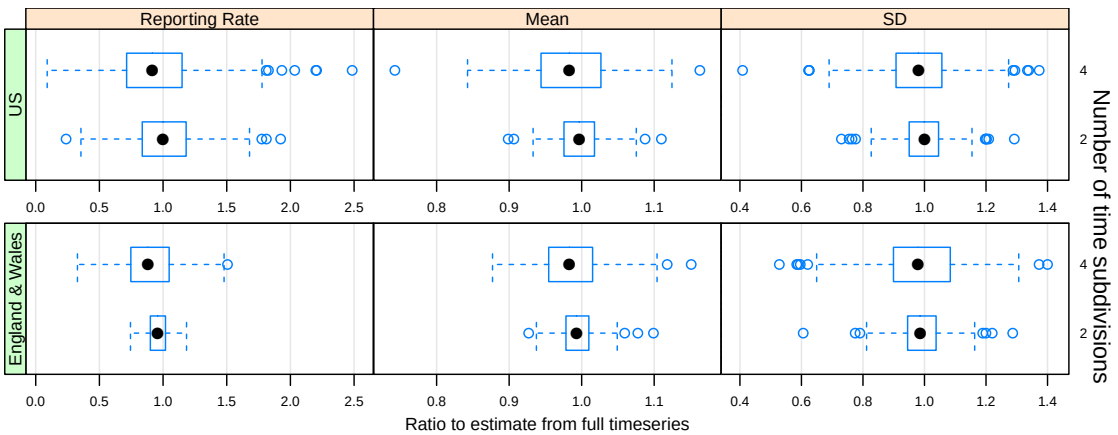

Figure 2.12. Ratio of inferred estimates from N_t equal-length subdivisions of timeseries to that of full timeseries ($\sim$ 20 years for each country), for N_t = 2 or 4. Mean = $\hat{\mu}$ and SD = $\hat{\sigma}$. R_0 is held constant at 20. Reporting rate is the estimate most sensitive to time length, particularly in the U.S. For $N_t = 2$, the inter-quartile range (IQR) of estimated reporting rates is within $\pm 20\%$ of the full length value for the U.S., and within $\pm 10\%$ in England & Wales. This suggests that reporting rate is relatively conserved over time for many populations. For $N_t > 4$ (timeseries less than 5 years), estimates diverge significantly.

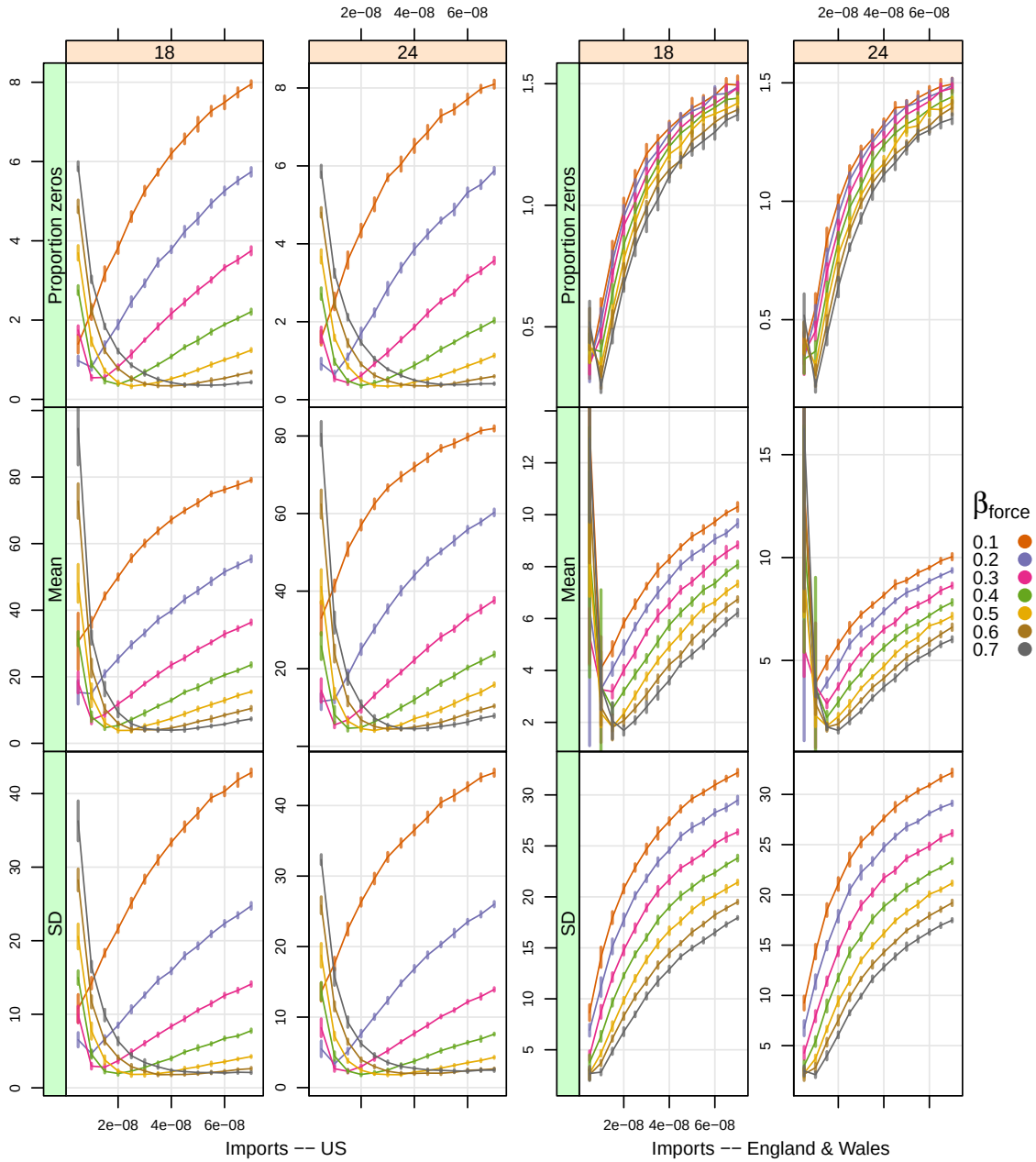

Figure 2.13. An ensemble of 10 realizations was constructed for each parameter set (columns show R_0; see Table 2.4 for tested parameter ranges). The residual sum of squares (RSS_δ) between model and data was computed for each realization and measure $\delta \in \{P_0, \hat{\mu}, \hat{\sigma}\}$. The ensemble mean RSS_δ is shown above, with vertical bars showing $\pm$ one standard deviation. For each δ, the condition with minimum ensemble mean RSS was identified. A t-test was conducted between the minimum set of RSS and each other condition's RSS set. Any parameter set that was greater than the minimum set with $p < 0.05$ was considered inferior to the best set for that measure, with the remaining sets considered equivalent. A final parameter set was chosen to maximize the number of measures in which it was equivalent to the best condition.

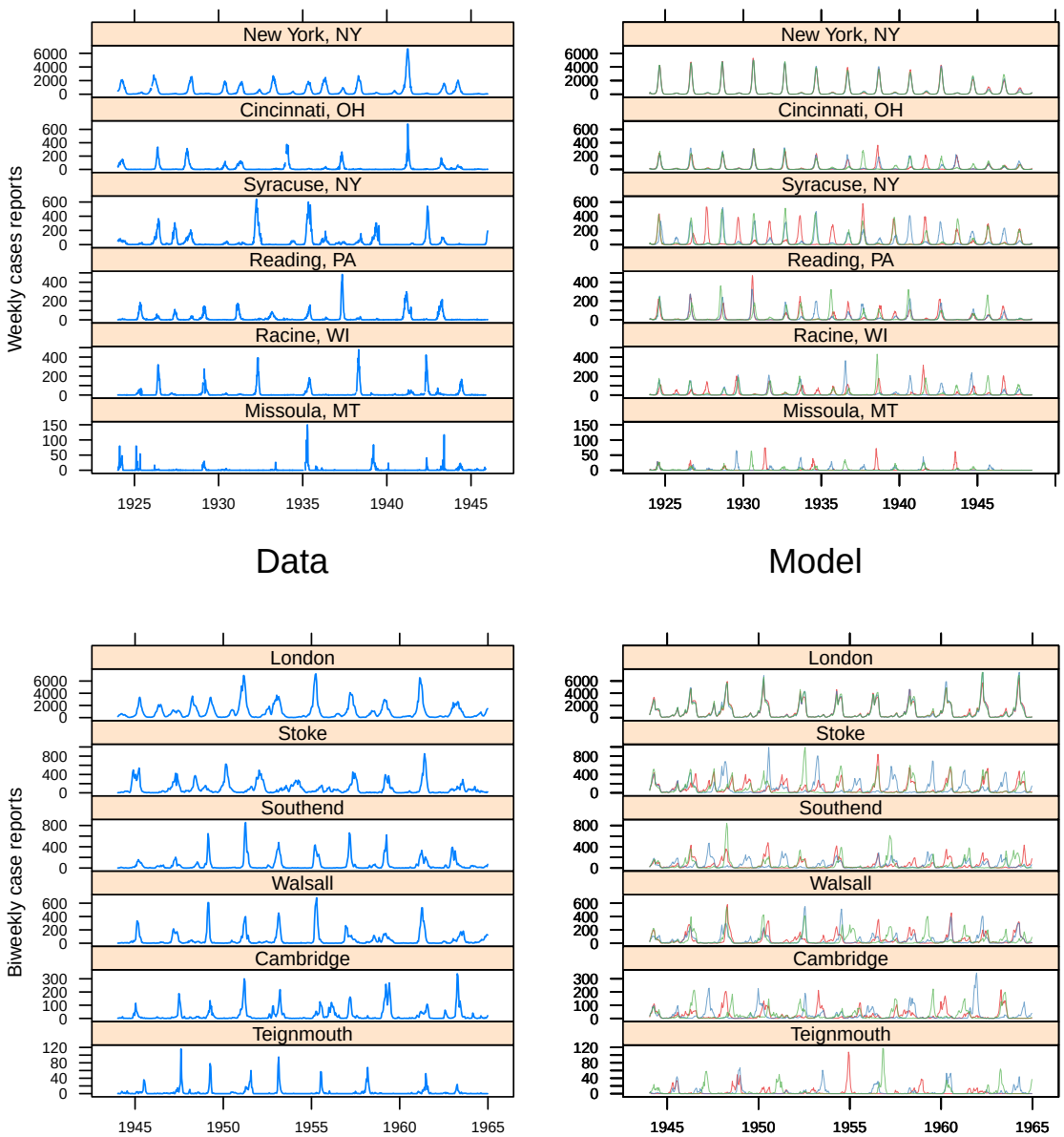

Figure 2.14. Simulation/data comparison of timeseries for select cities. Cities are ordered by descending population, with evenly-spaced population ranks. Data is shown on the left, and three randomly-selected simulations are shown on the right. To simplify comparison, matching cities share a common Y axis. For model parameters, see Table 2.5.

2.8 Appendix S2 – Epidemiological Model Details

Event	Change in State	Transition Rate
Birth (discounting infant mortality)	$(S, E, I, R) \to (S+1, E, I, R)$	$\nu(t)N$
Immigration of susceptible	$(S, E, I, R) \to (S+1, E, I, R)$	$\mu_+(t)(\frac{1}{R_0})N$
Emigration of susceptible	$(S, E, I, R) \to (S-1, E, I, R)$	$\mu_-(t)(\frac{1}{R_0})N$
Immigration of recovered	$(S, E, I, R) \to (S, E, I, R+1)$	$\mu_+(t)(1-\frac{1}{R_0})N$
Death or emigration of recovered	$(S, E, I, R) \to (S, E, I, R-1)$	$\mu_-(t)(1-\frac{1}{R_0})N$
Exposure due to imports	$(S, E, I, R) \to (S-1, E+1, I, R)$	$\eta\hat{\beta}S\frac{\sum_{j\neq i} I_j}{\sum_j N_j}$
Exposure due to internal dynamics	$(S, E, I, R) \to (S-1, E+1, I, R)$	$\hat{\beta}S\frac{I}{N}$
Infection	$(S, E, I, R) \to (S, E-1, I+1, R)$	σE
Recovery	$(S, E, I, R) \to (S, E, I-1, R+1)$	γI

Table 2.3. Events and corresponding transition rates in the stochastic $SEIR$ model for population i.

Parameter	Range
$\nu(t)$	Specified by demographic data
$\mu(t)$	Specified by demographic data
η	$[10^{-9}, 10^{-6}]$
$R_0 = \frac{\beta_0}{\gamma}$	$[15, 25]$
β_{force}	$[0.1, 0.5]$
$\hat{\beta}$	$\beta_0(1+\beta_1 \sin(\frac{2\pi t}{365}))$ per day
σ	1/8 per day
γ	1/5 per day

Table 2.4. Parameter values.

Each model realization was initialized at the equilibrium values of the equivalent non-seasonal deterministic model and run over years where demographics were available; the U.S. model was run from 1910 through 1949, and the England & Wales model was run from 1944 through 1964. Model results outside the observed time series range were discarded.

Country	R_0	β_{force}	η	School Forcing Function
US	18	0.50	2.50e-08	Sin
England & Wales	24	0.70	1.00e-08	Term Time

Table 2.5. Table of final epidemiological model parameters. School term in England & Wales as per Keeling et al. [11].

We tested a number of model innovations including term-time forcing, seasonal forcing of η, and scaling η with population size. We tested each model innovation over a range of parameter values. Specifically, we varied the transmission rate, strength of seasonal forcing, size of the import term, and R_0 in turn while holding all other parameters constant. For further analysis, we selected the simplest possible model structure as it yielded results comparable to more complex models.

Chapter 3

Reporting Rate Variability

Incomplete reporting of pre-vaccine era childhood diseases: a case study of observation process variability

Christian Gunning[1,*] Erik Erhardt[2], Helen J. Wearing[1,2],

1 Department of Biology, University of New Mexico, Albuquerque, New Mexico, USA

2 Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA

∗ **E-mail: xian@unm.edu**

3.1 Abstract

Incomplete observation is an important yet often neglected feature of historical ecological timeseries. In particular, historical case report timeseries of childhood diseases have played an important role in the formulation of mechanistic dynamical models of populations and metapopulations. Yet no comprehensive study of

childhood disease reporting probabilities (commonly referred to as reporting rate) has been conducted to date. Here we provide a detailed analysis of measles and whooping cough reporting probabilities in pre-vaccine U.S. cities and states, as well as cities in England & Wales. Overall, we find the variability between locations and diseases greatly exceeds that between methods or time periods. We demonstrate a strong relationship within location between diseases, and within disease between geographic areas. In addition, we find that demographic covariates such as ethnic composition and school attendance explain a nontrivial proportion of reporting probability variation. Overall, our findings show that disease reporting is both variable and non-random, and that completeness of reporting is influenced by disease identity, geography, and socioeconomic factors. We suggest that variability in observation processes such as incomplete reporting an be accounted for, and that doing so can reveal key dynamical processes that are otherwise obscured.

3.2 Introduction

Historical datasets have long aided ecologists in unraveling the complex dynamical interactions of real-world populations and metapopulations. In particular, observational datasets can provide extensive spatial and temporal coverage difficult to achieve through field experiments. From disease ecology to wildlife and natural resource ecology, these datasets have allowed ecologists to evaluate the strength and significance of a wide range of dynamical processes [1–11].

Historical datasets are shaped by a variety of observation processes. Though not the core focus of ecological interest, imperfect observation is a rule rather than an exception in datasets resulting from surveillance rather than controlled experimentation. The extent to which imperfect observation such as incomplete and variable disease reporting can distort or obscure dynamical processes such as local extinction

remains an open question, as does the ability to correct for imperfect observation. When observation processes are stationary and independent of mechanistic dynamical processes, post hoc estimation of the state variables of interest is possible using known constraints of the dynamical system.

The study of human infectious diseases has yielded important insights into the non-linear dynamics of real-world populations and metapopulations, largely due to extensive historical datasets. For human diseases such as measles, cities comprise the basic epidemiological units of observation over which disease dynamics (and reporting probabilities) are typically assumed to be constant. Reporting of human infectious diseases is known to be both imperfect and variable between cities [12–16]. A reporting probability (proportion of true infections recorded as official case reports) can be estimated for acute, highly-infectious diseases that confer permanent immunity, using a combination of demographic and case report data. Here, we show that reporting probabilities of human infectious diseases follow conserved patterns in space and time. By accounting for reporting probabilities, we also provide a more accurate estimate of the scaling of local persistence with population size.

Reporting probabilities of childhood diseases received considerable attention throughout the 20^{th} century in both the U.S. [14] and England & Wales [15]. Notable works include Bartlett [2], who reviews estimates from the early 20^{th} century in both countries, Black [17], who reports summary estimates for several countries, as well as Finkenstädt and Grenfell [5] and Bjørnstad et al. [18], who employ the susceptible reconstruction method. However, we have found no systematic review of variation in childhood diseases reporting probability between populations (cities) and metapopulations (here, countries).

Stochastic extinction within a host population such as a city is driven by local processes, such as host demographics, and metapopulation processes, such as disease importation between populations. Yet stochastic extinction is not easily distin-

guished from incomplete reporting. Several works have explicitly incorporated estimation of incomplete and variable reporting into dynamical models of populations [8, 9] and metapopulations [11]. Nonetheless, disease reporting (and variability thereof) has been largely absent from modern population and metapopulation models that studied stochastic extinction and disease persistence in England & Wales [19–22]. These models (and results) do not necessarily generalize to metapopulations with lower and more variable reporting, such as the pre-vaccine U.S. or modern sub-Saharan Africa.

Outline

This study aims to quantify and explain variability in the reporting probabilities of two childhood diseases prior to mass vaccination. We use an extensive dataset of measles [11] and whooping cough case reports in U.S. states and cities in the pre-vaccine era, in addition to the classic 60-city England & Wales measles dataset [18]. To estimate the total per-population susceptible pool, we employ two different sources of demographic records. Using case reports and susceptibles, we then compute the reporting probability of each disease and location (cities or states).

Here we refer to sampled units (e.g. specific cities and states) as *locations*, while *area* refers to the level of geographic sampling (i.e., city versus state). For human diseases such as measles and whooping cough, each city is a coherent epidemiological population, throughout which disease dynamics (and reporting probabilities) are typically assumed to be constant. U.S. states, on the other hand, are primarily administrative subdivisions that are socially and epidemiologically heterogeneous. Thus, state reporting probability estimates are assumed to be averaged over many discrete populations (e.g., cities and towns). Nonetheless, the unambiguous nesting of cities within states provides a useful estimate of the effect of geographic location.

We begin with a comparison of reporting probabilities between diseases and between geographic areas (e.g., between states and their respective cities). We find a very strong relationship within location between diseases, and a strong relationship within disease between geographic areas. In addition, we explore the temporal variation of reporting probability in cities. The strong dependence of reporting probability on geographical identity rather than time suggests that socioeconomic factors strongly influence disease reporting probability. Indeed, we find that a nontrivial proportion of variation in reporting probability is explained by the proportion of a location's population that is either white or attending school.

We also include a discussion of uncertainty and sources of error. Metadata detailing the collection process of both case reports and demographic records is often sparse or altogether lacking. Here, we use several independent sources of demographic data, two different methods of calculation (per capita rates and census microdata), and bootstrap estimates for one method. Overall, we find the variability between locations and diseases greatly exceeds that between methods or time periods.

We conclude with a discussion of metapopulation dynamics and the obscuring effects of incomplete reporting. In observational datasets, poor reporting is indistinguishable from stochastic extinction in individual populations (e.g. cities). Correcting for variable reporting regenerates the hypothesized scaling relationship between population size and observed extinction in the studied metapopulations.

3.3 Results

Overall, a high degree of variability in disease reporting was observed between both locations and diseases. The distribution of reporting probabilities for each area (cities, states) and disease (using demographic method results, available for all locations) is shown in Figure 3.1, and summary statistics are shown in Table 3.1.

In the U.S., whooping cough probabilities are much lower than measles probabilities, regardless of area. The cities of England & Wales have higher and less variable measles reporting probabilities than U.S. cities or states, consistent with previous estimates [2, 14, 15].

Comparison between time periods

The employed method assumes that each system is approximately stationary over the period of study, and the systems in question should be assessed for major perturbations during the period of study. In addition, a sufficiently long period of time must be employed such that stochastic and season fluctuations are short relative to the full period of record.

In order to assess temporal variation in reporting probabilities in the present systems, city case reports were subdivided into two time series of approximately equal length (Table 3.3). Figure 3.2 shows that city reporting probability is relatively invariant across time, though more temporal variation is evident in England & Wales. The National Health Service was fully implemented in the United Kingdom by 1948. This change in public health infrastructure could explain some of the observed temporal variation in England & Wales (Figure 3.2), though any metapopulation-level temporal shift is slight.

Comparison between methods

Case report totals for each disease are identical between methods, with different reporting probability estimates (Figure 3.5 and 3.6) resulting from variations in each location's total susceptible pool. For U.S. states, estimated reporting probabilities were highly conserved between methods: between-method linear models yield a

slope $\sim$ unity and a non-significant intercept. For U.S. cities, reporting probabilities estimated from the census method are slightly lower than those from the per capita method.

One major limitation of the per capita method is that U.S. city birth rates are inferred from the per capita rates of their respective states. In general, states have higher birth rates than cities in this era (Figure 3.7) This is likely due to states' rural populations, which have generally higher birth rates than urban areas in this era [23]. Consequently, the census method likely over-estimates U.S. city susceptibles and under-estimates reporting probabilities, as observed here.

For reference, the per capita method was also compared with the susceptible reconstruction method [5] for all areas (see Methods for important assumptions). Susceptible reconstruction yielded slightly higher estimates, particularly in England & Wales, though the differences are small (Figure 3.8).

Conserved patterns of variation in disease reporting

The interdependence of disease identity and geographic location in the U.S. is shown in Figure 3.3. Reporting probabilities of whooping cough are strongly correlated with measles probabilities, regardless of area (Figure 3.3A). Though more scatter is evident, city reporting probabilities are correlated with their associated states' probabilities, regardless of disease (Figure 3.3B). Estimated slopes and correlation coefficients for each linear model specification, along with bootstrapped confidence intervals, are shown in Table 3.2. Overall, we find that the observation process of disease reporting is conserved over space and time, and that disease identity and geography influences reporting probabilities in consistent ways.

Within-location variability estimates derived from bootstrapping of census micro-data are shown in Figure 3.3 (reporting probabilities and confidence intervals are

shown in Tables 3.4 and 3.5). Bootstrap estimates show that larger locations consistently exhibit less variation, as expected (Figure 3.9). Overall, between-location variation greatly exceeds within-location variation, increasing confidence in the observed patterns.

The influence of socioeconomic identity on incomplete reporting was explicitly modeled (Table 3.6). A range of demographic covariates were tested using forward model selection (see Methods). While many of these predictors are correlated, forward model selection favors a parsimonious model by selecting the best predictors first, as shown in Table 3.6. The final models explain much of the observed variation in reporting probabilities: $R^2 = 0.51$ (state measles); $R^2 = 0.4$ (state whooping cough); $R^2 = 0.32$ (city measles); $R^2 = 0.13$ (city whooping cough). Overall, variation in measles reporting probabilities is much better explained by demographic covariates than that of whooping cough.

Two covariates emerged as most significant: the proportion of a location's population that is either white (prop.white) or attending school (prop.school). Regardless of disease, higher reporting probabilities are correlated with a higher proportion white for states and a higher proportion attending school for cities. Other significant predictors include proportion in labor force (states, both diseases, positive correlation), household size SD (states, both diseases, positive correlation), proportion male (states, whooping cough, negative correlation), and mean household size (cities, measles, negative correlation). Overall, selected covariates and their associated parameter estimates are generally consistent between diseases within each area.

Causal mechanisms of the observed correlations remains unclear. Nonetheless, the significant covariates broadly relate to measures of economic status (ethnic composition, labor force, and sex ratio) as well as indicators of social structures that can influence the distribution infection age and disease reporting (household size distribution, schooling).

3.4 Discussion

Historical datasets are valuable for their wide spatiotemporal extent, yet their post-hoc observational nature means that key dynamical processes such as stochastic extinction can be obscured by imperfect and variable observation. Measles and whooping cough are two well-studied childhood diseases with very different symptomology, epidemiology, and temporal dynamics. Yet both infect the majority of susceptible individuals in childhood and confer lasting immunity. In addition, both diseases undergo stochastic extinction at a rate dependent on population size and birth rate [24]. Here we infer a key observation process using a long-term constraint of each dynamical system, i.e. the mass balance of susceptibles in childhood diseases. We find that reporting probabilities vary greatly between disease, geographic region, and metapopulation. This variability directly affects patterns of observed extinctions or "fade-outs" [2, 22, 25] and, if not addressed, makes comparisons between diseases and metapopulations difficult.

We find that measles reporting probabilities of cities in the U.S. are lower and more variable than in England & Wales. In the U.S., we find that measles is better reported than whooping cough (as previously found in England & Wales by Clarkson and Fine [15]). In addition, we find that reporting probability varies consistently by geographic locale: those locations that report measles well also report whooping cough well, and vice versa. On the other hand, reporting probabilities do not appear to vary appreciably by time in either country or disease in the eras considered. Likewise, bootstrapping indicates that between-location variation greatly exceeds within-location uncertainty in the U.S. Finally, we show that demographic covariates, including proportion white and proportion attending school, explain a non-trivial proportion of the observed variation in U.S. reporting probabilities: locations that have low school attendance and high minority populations have lower reporting probabilities, regardless of disease. Overall, we find substantial spatial, temporal,

and socioeconomic consistency within the pronounced heterogeneity of pre-vaccine era disease reporting.

Estimating uncertainty

Historical datasets frequently lack detailed metadata, including full descriptions of sampling protocols. This introduces a persistent difficulty of estimating uncertainty and establishing concordance between varying data sources. For example, we find here that the geographic definition of certain cities is not comparable between census microdata and case reports. Additionally, we have no detailed definition of the geographic limits used to define cities in case report collections. In short, we cannot unambiguously identify all sources of error and uncertainty. Nonetheless, we can often constrain error and uncertainty, for example by the comparison of multiple, independent data sources, as we do in this study with demographic data.

The structure of census microdata allows us to estimate the sampling distribution of each location's susceptible population by bootstrapping each decadal census. The results clearly show that observed variation decreases for larger locations (Figure 3.9). These variance estimates do not account for the processes that generate case reports. Here, we assume that the reporting probability fully describes these processes, and is time invariant.

Migration remains an interesting problem deserving further attention, and can possibly be estimated at the decadal level from census microdata. Here we assume a minimal impact of migration. In the U.S., the overall flow of migration is to urban areas from rural areas, which we expect experience much higher levels of stochastic extinction and consequently higher and more variable ages of infection (for an extreme example, see discussion by Crum [26] of U.S. Civil War troops). Large rural-to-urban migration waves, as seen in the Great Migration [27], could

result in an underestimate of susceptibles and overestimate in reporting probability for some cities. Nonetheless, the low average age of infection of both measles and whooping cough in the pre-vaccine U.S. and England & Wales [28] suggests that the vast majority of migrants are not susceptible to either disease.

Importance of observation processes in dynamical process estimation

As noted above, poor disease reporting and stochastic extinction cannot be easily separated, particularly in cities that regularly teeter on the boundary of extinction [11]. This conflation means that metapopulation-level scaling patterns of stochastic extinction are complicated by between-population variation in incomplete reporting. In this regard, the large body of work on measles in England & Wales that neglected reporting probabilities [19, 21, 29] has benefited from the happy accident of relatively high and uniform reporting probabilities. In the U.S., on the other hand, failing to account for the low and variable disease reporting in this era paints a false picture of the overall metapopulation dynamics and hinders a comparison between metapopulations [11].

To date, a number of ad hoc measures have been employed to address incomplete disease reporting. The proportion of zero observations (over a suitably long period of time) is one common measure of stochastic extinction [8, 11, 18, 30, 31]. This measure is appealing due to its simplicity, but has been criticized as sensitive to disease reporting. In an attempt to address these concerns [2] employed a 3-week period of observed extinction, termed fade-out. Conlan et al. [22] propose several alternate measures, including fade-outs post invasion and fade-outs post epidemic. These methods were proposed on mechanistic grounds, yet their respective frequencies of false negatives (apparent extinction) and true positives are not well-characterized

under the low and variable disease reporting. In addition, several measures (e.g., Conlan et al. [22]) depend on a priori threshold values.

The proportion of observed zeros and its scaling with population size provides a useful example in the present systems. Previous work has demonstrated a strong interaction between proportion of zero observations and poor disease reporting in the U.S., particularly in medium-sized cities that hover at the edge of extinction Gunning and Wearing [11]. Under the assumption of homogeneous mixing, the reporting probability is equivalent to the proportion of the total population under surveillance (homogeneous mixing is a poor assumption for U.S. states, which are not considered). Here we demonstrate a simple rescaling of total population size by reporting probability to yield an effective population size. The result is log-linear dependence of extinction risk on effective population size (Figure 3.4). While the longer 2 week sampling period of England & Wales also affects the probability of zero-observations, this rescaling produces a concordance between the two countries, similar to the pattern observed in Gunning and Wearing [11]. Overall, this example illustrates the importance and relative ease of accounting for variation in observation processes when characterizing real-world dynamical systems.

A comparison between the U.S. and England & Wales also highlights the role of socioeconomic diversity. Our results (Table 3.6) suggest that high levels of ethnic and cultural heterogeneity, as seen in the U.S. compared to pre-vaccine England & Wales, increases variation in disease reporting. Indeed, less complete reporting in U.S. minority populations was suggested a century ago by Crum [26]. This pattern warrants testing in the modern era in regions such as Niger, which have large rural populations and a small number of large cities [32]. In a socioeconomically heterogeneous state such as Niger, significant variation in measles reporting probabilities appears to be a conservative assumption. Furthermore, observed variation in reporting of other human infectious diseases can be explained by similar socioeconomic

disparities. For example, Undurraga et al. [16] showed that the estimated probability of underreporting of dengue episodes at a national level in Southeast Asia and the Americas correlated with a measure of health quality.

Broader applications

The employed method is only appropriate for fully immunizing diseases. In the modern era, vaccination introduces additional sources of variation and measurement error, since estimates of both vaccine uptake and efficacy are required [15]. In addition, we generate single estimates from long time periods (e.g., multiple epidemics), and assume minimal temporal variation in disease reporting.

Nonetheless, we propose that disease ecologists and epidemiologists can often estimate between-population variation in observation processes. Accounting for this variation appears to be particularly important in socioeconomically diverse populations. The framework that we employ is conceptually and analytically simple, provided sufficient demographic information is available. Indeed, even when relevant demographic details are sparse or absent, rough estimates of reporting probability can suggest whether or not between-location variation overwhelms the dynamical processes of interest.

3.5 Materials and Methods

Case reports

U.S. weekly case reports of measles and whooping cough were obtained as PDFs from the United States Public Health Reports [33]. Case reports were manually double-entered using a custom "Mechanical Turk" web application that automati-

cally identified conflicts, which were manually resolved. Populations that contained more than 20% missing values for either disease were removed. Populations were also removed if demographic data was unavailable (see below). In later time periods in the U.S., sample coverage grows sporadic, particularly for cities. To avoid bias from temporally aggregated missing data, years were excluded if more than 50% of the remaining cities had fewer than 45 sampled weeks. Missing case reports were excluded from further analysis.

Measles case reports in England & Wales were originally recorded by the United Kingdom Office of Population Censuses and Surveys [34, 35]. We employ the publically available 60-city subset studied by Bjørnstad et al. [18]. This case report dataset was downsampled to a 2 week sampling interval (this is twice the sampling period of the U.S., though the difference has no effect on reporting probability calculations). City-level case reports of whooping cough in England & Wales have been studied extensively [31, 36], but are not publicly available at this time.

Case report lengths and boundaries are shown in Table 3.1, and plotted in Figures 3.10-3.14. In the U.S., 48 cities and 46 states were selected for final analysis, as well as 60 cities from England & Wales.

Demographic data

For U.S. locations (cities and states), two main sources of demographic data were used to estimate the total susceptible pool over the period of case report records.

Per capita method

For the per capita method, each location's total susceptible pool was obtained from yearly population estimates and per capita birth, death, and infant mortality rates.

First, decadal populations were obtained from the U.S. decadal census (1920-1950) [37]. Yearly populations were estimated using an exponential growth model to interpolate between decadal population. Yearly state per capita birth and death rates were obtained from the U.S. National Center for Health Statistics [38, 39]. Yearly per capita national infant mortality rates were obtained from the U.S. Census Bureau [40].

Yearly populations were used to calculate the flow of new susceptibles into each location from state birth rates (discounted by national infant mortality rates). Finally, susceptible flows were summed over the period of record of each disease and location to infer the total susceptible pool. For this method, pre-infection migration and (non-infant) death of susceptibles was assumed to be minimal.

Microdata method

Census microdata refers to the original responses of each individual in a country's census, and includes a range of variables for each response such as location, age, gender, and ethnicity. These data are currently only available for the U.S. decadal census, and were obtained from the Integrated Public Use Microdata Series (IPUMS, 1920-1950, 1% sample) [41]. Due to privacy concerns, no microdata is available for cities with populations of less than 25,000 as of 1920. In addition, the census boundaries of several cities changed over time. Cities falling in either of these groups were excluded from further analysis.

Census microdata was used to estimate the total susceptible pool of each U.S. location. Individuals of ages 1 through 10 (inclusive) that were born within the period of record of each disease and location were summed to infer the total susceptible pool. This method includes all migration and death of susceptibles of age ≤ 10 years at the time of census that occurred in the decade preceding the census.

Census microdata permits an assessment of the sampling distribution of the susceptible pool via bootstrapping. Each census was bootstrapped 1E+04 times for each disease, and the total susceptible pool recomputed for each bootstrap.

Comparability

For several reasons, these methods are not strictly comparable. First, the census microdata of several cities expands to include neighboring cities in 1940 and 1950 (such as Tampa and St. Petersburg, FL and Minneapolis and St. Paul, MN), leading to a detectable overestimate of susceptibles in select cities. These cities were excluded from further analysis. Second, state birth rates are used in the per capita method to estimate births of both states and their associated cities (see Discussion). Finally, yearly city births for England & Wales were provided by Rohani [42], and were subsequently adjusted by the national infant mortality rate. This method is closest to the per capita method.

Reporting Probability

We begin by assuming that reporting probabilities (commonly referred to as reporting rates) are invariant over time within each disease and location. For each location i and disease j, we have obtained the total number of observed case reports C_{ij} and total (new) susceptible pool S_{ij} (note that, since the observation window varies between disease, S_i depends upon disease). If the epidemiological system is approximately stationary over the time period considered (i.e., there are no major changes in the underlying processes governing the disease and demographic dynamics), then the number of susceptible individuals in the population should also be approximately stationary. This implies that the flow of new susceptibles is counterbalanced by the flow of new infections. For a disease that confers permanent immunity, new suscep-

tibles are just surviving births (ignoring the effects of migration). The simplest estimate of reporting probability is therefore obtained by assuming that the total number of expected cases, E_{ij}, is approximately equal to the total accumulated susceptible pool, S_{ij}, over the period of interest. Thus, the reporting probability $r_{ij} = \frac{C_{ij}}{E_{ij}} \approx \frac{C_{ij}}{S_{ij}}$ [15].

This simple estimate assumes that the number of individuals who are susceptible at the beginning of the time period considered is approximately the same as the number susceptible at the end. Previous work [5] has regressed cumulative births against cumulative cases, and obtained an estimate of reporting probability as the slope of the regression line. The two estimates are the same if the deviation from the average number of susceptibles is the same at the beginning and end of the time period. This can be achieved in a stationary system if the time period considered begins and ends at approximately the same point in the epidemic cycle.

To assess the time variability of reporting probabilities, we subdivide U.S. city case reports into two approximately equal subdivisions (Early and Late) and re-estimate reporting probability using the per capita method. We also estimate reporting probabilities using the susceptible reconstruction method [5], where the reporting probability is the slope of the linear regression between cumulative yearly births and case reports. Notably, we aggregate to a yearly scale to conduct susceptible reconstruction, since we have no knowledge of within-year variability in birth rates.

Modeling the interdependence of reporting probabilities

Each location (city, state) has an associated reporting probability for both measles and whooping cough (between-disease). Likewise, each disease has an associated reporting probability for each state and its associated cities (between-area). We employ a set of linear models to quantify the interdependence in between-disease

and between-area reporting probabilities.

For the between-disease case, we model whooping cough as a function of measles, with a separate model for each area. For the between-area case, we model cities as a function of their associated states, with a separate model for each disease. The result is four separate model specifications. Our choice of measles as the independent variable is arbitrary. On the other hand, cities have a natural dependence on states due to the nested nature of public health administration. For simplicity, we use ordinary least-squares (OLS) regression rather than error-in-variables regression. All reporting probabilities were logit_2-transformed to correct for heteroskedasticity. The logit_2-transform is simply $log_2(\frac{p}{1-p})$, such that one unit of increase equates with a doubling the reporting probability odds, e.g., from 50% (1/1 odds, $\text{logit}_2(\text{odds})=0$) to 66% (2/1 odds, $\text{logit}_2(\text{odds})=1$).

For each linear model specification, 1E+04 model realizations were constructed via bootstrap resampling. For each realization, a two-step sampling process was employed. First, city identity was sampled with replacement. Second, for each sampled city, the reporting probability of the relevant independent and dependent variables were sampled with replacement from the appropriate bootstrap distribution. Finally, simple linear regression was conducted on the resulting sample, and the slope, intercept, and correlation coefficient were extracted. This strategy, known as "bootstrapping pairs" [43], accounts for uncertainty in reporting probabilities both within and between cities without making standard normality and constant-variance assumptions on the residuals. Rather, this strategy assumes only that the cities are randomly sampled from the population distribution of cities (see above for city selection criteria).

Demographic covariate models of reporting probabilities

Census microdata provides a number of demographic covariates at both the individual and household level for each location and census. Due to complete coverage of sampled locations, the 1930 decadal census was selected for further analysis. A weighted summary of each covariate was calculated by location to yield either a proportion (categorical variables) or mean and standard deviation (continuous variables). A separate linear model was then constructed for each disease and area to model reporting probability (microdata method) as a function of covariate summaries. Each model was constructed via forward selection with a BIC selection criteria. In addition, reporting probability was logit_2 transformed and all predictors were zero-centered prior to model construction.

Tested covariates include the proportion of each location's population that was white (prop.white), in school (prop.school), male (prop.male), born in the current state of residence (prop.local), and in the labor force (prop.labforce). In addition, the mean and standard deviation of age (mean.age and sd.age) and household size (mean.housesize and sd.housesize) was also tested. Note that covariate summaries are broadly correlated.

Acknowledgments

The authors would like to thank Natalie Wright, Robert Liberatore, Nicholas Giron, Pej Rohani, and Matthew Ferrari for their assistance. Joe Conway assisted with database design.

CG was supported by a fellowship in the Program in Interdisciplinary Biological and Biomedical Sciences at the University of New Mexico. This publication was made possible by Grant Numbers P20RR018754 from the National Center for

Research Resources (NCRR), T32EB009414 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), components of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR, NIBIB, NIGMS or NIH.

Data Accessibility

- U.S. case report data and demographics: See Data Dryad
- England & Wales case reports: Data from `http://www.zoo.cam.ac.uk/zoostaff/grenfell/measles.htm`, no longer available.
- England & Wales demographics: Personal communication with Dr. Pej Rohani (rohani@umich.edu.
- Integrated Public Use Microdata Series: `https://usa.ipums.org/usa/`

3.6 Figures

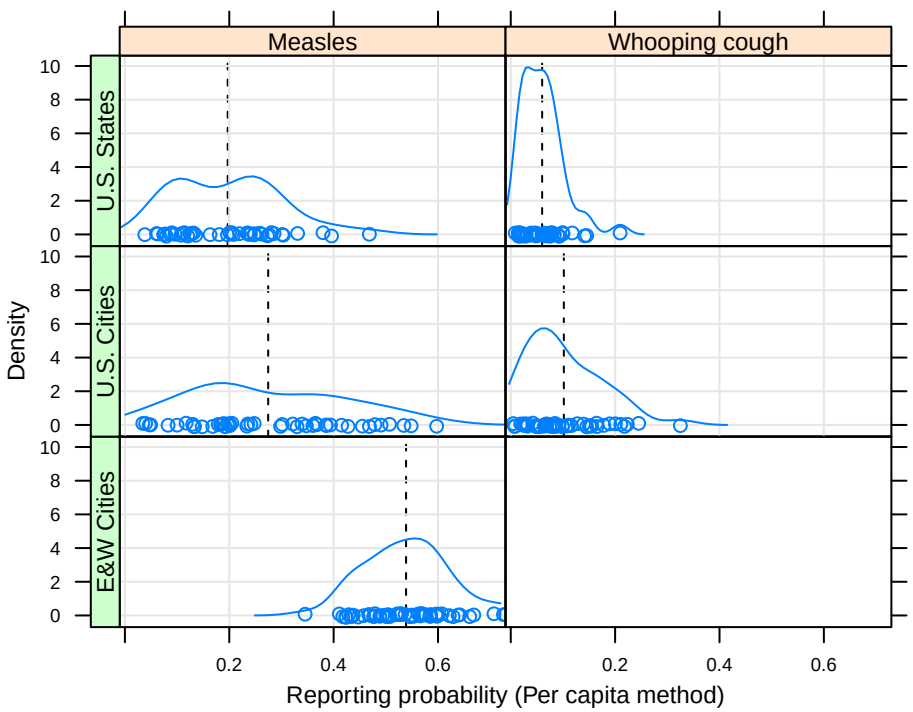

Figure 3.1. Distribution of reporting probability estimates for all diseases and areas. Demographic method results (shown here) are available for all localities. Overall, reporting of whooping cough is less complete than measles, while U.S. reporting is less complete than England & Wales (E&W). Extensive variation between locations is evident, particularly in the U.S. Black dashed line: group mean.

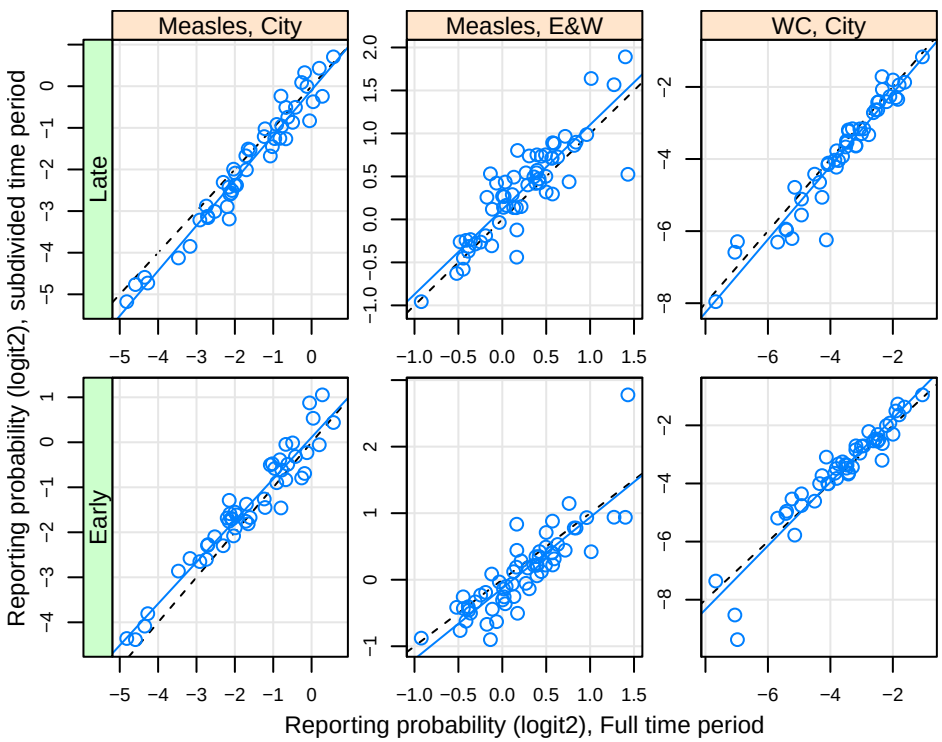

Figure 3.2. Time period comparison (per capita method, E&W = England & Wales, WC = whooping cough). For cities, case report timeseries were divided into early and late portions (see Table 3.3). Reporting probability estimated from these sub-periods generally match those from the full period. U.S. city measles is the exception, with more complete reporting in the early period. All values were logit_2 transformed to correct for heteroskedasticity. Black dashed line: 1-1 line; solid blue line: linear model.

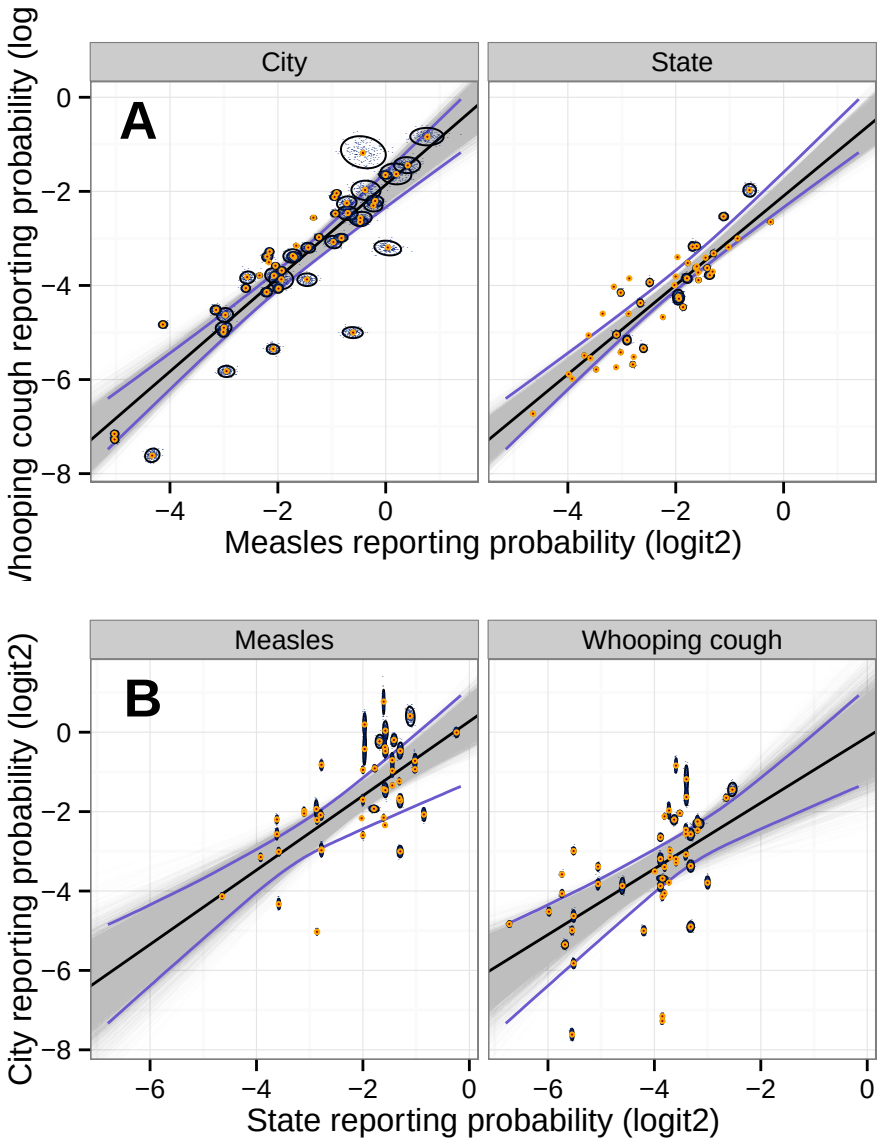

Figure 3.3. Reporting probability estimates, showing that the variability between locations (cities or states) greatly exceeds variability within locations. 1E+04 total bootstraps were drawn (200 are plotted, small black points). For each location, an approximate 95% confidence interval (black ovals) and median probability (orange central dot) are shown. The median linear model (black line) and approximate model 95% CI (blue lines) are also plotted. Linear model results and bootstrap confidence intervals are shown in Table 3.2.
A Whooping cough reporting probabilities are closely proportional to those of measles, regardless of area.
B City reporting probabilities are roughly proportional to those of the associated states, regardless of disease.

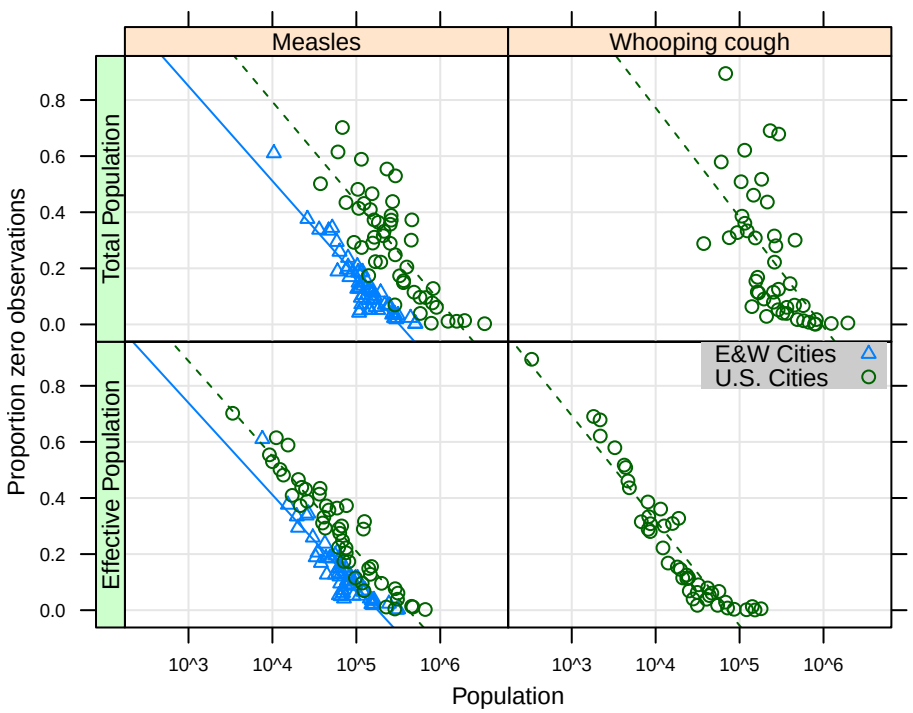

Figure 3.4. Semi-log scaling of observed zeros with total population (TP) and effective population (EP = TP * reporting probability). Proportion zeros shows a closer scaling with EP than TP, particularly in whooping cough, as well as a closer correspondence between metapopulations. Regressions for each metapopulation are overplotted. Populations with no observed zeros are excluded. The 2 week sampling interval of England & Wales (E&W) is twice that of the U.S. (1 week). For the EP population model, $R^2 = 0.89$ (U.S. measles), 0.82 (E&W measles), 0.90 (U.S. whooping cough).

3.7 Tables

Disease	Area	L	N	Start	End	Mean	SD
Measles	U.S. Cities	48	1148	1924-01-05	1945-12-29	0.27	0.15
Measles	E&W Cities	60	598	1944-01-09	1966-12-25	0.54	0.08
Measles	U.S. States	45	1089	1928-01-07	1948-12-11	0.20	0.10
WC	U.S. Cities	48	1148	1924-01-05	1945-12-29	0.10	0.07
WC	U.S. States	46	467	1938-01-01	1946-12-07	0.06	0.04

Table 3.1. Number of sampled locations (L), number of sampled case reports (N), and time range, as well as summary statistics of estimated reporting probabilities (per capita method) for each disease and area. Note the limited sample coverage for state whooping cough reports. U.S. locations are sampled weekly; England & Wales cities are sampled every other week. WC: whooping cough; E&W: England & Wales.

Model	Subset	Slope	CI	Correlation	CI
City vs. State	Measles	0.93	(0.67, 1.23)	0.66	(0.50, 0.77)
City vs. State	Whooping cough	0.82	(0.53, 1.19)	0.55	(0.34, 0.73)
Measles vs. Whooping cough	City	1.00	(0.81, 1.15)	0.90	(0.78, 0.96)
Measles vs. Whooping cough	State	0.95	(0.82, 1.07)	0.89	(0.82, 0.94)

Table 3.2. Linear models of reporting probability variation (microdata method) between area and between disease, constructed from 1E+04 bootstrap draws (see Methods). Median and 95% CI of model results and correlations are shown. Slope estimates are in logit_2 units. For the measles subset, doubling the state reporting probability odds, i.e. from 50% (1/1 odds, $\text{logit}_2(\text{odds})=0$) to 66% (2/1 odds, $\text{logit}_2(\text{odds})=1$) approximately doubles the city reporting probability odds.

3.8 References

[1] C. Elton and Mary N. The ten-year cycle in numbers of the lynx in Canada. *J Anim Ecol*, 11(2):215–244, 1942.

[2] M.S. Bartlett. The critical community size for measles in the United States. *J R Stat Soc Ser A*, 123(1):37–44, 1960.

[3] D. Ludwig, D.D. Jones, and C.S. Holling. Qualitative Analysis of Insect Outbreak Systems: The Spruce Budworm and Forest. *J Anim Ecol*, 47(1): 315–332, 1978.

[4] A.A. Berryman. Can economic forces cause ecological chaos? The case of the northern California Dungeness crab fishery. *Oikos*, 62(1):106–109, 1991.

[5] B.F. Finkenstädt and B.T. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *J R Stat Soc Ser C Appl Stat*, 49(2):187–205, 2000.

[6] N.M. Ferguson, C.A. Donnelly, and R.M. Anderson. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science*, 292(5519):1155–1160, 2001.

[7] Y. Xia, O.N. Bjørnstad, and B.T. Grenfell. Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Nat.*, 164(2): 267–281, 2004.

[8] M.J. Ferrari, R.F. Grais, N. Bharti, A.J.K. Conlan, O.N. Bjørnstad, L.J. Wolfson, P.J. Guerin, A. Djibo, and B.T. Grenfell. The dynamics of measles in sub-Saharan Africa. *Nature*, 451(7179):679–684, 2008.

[9] D. He, E.L. Ionides, and A.A. King. Plug-and-play inference for disease dynam-

ics: measles in large and small populations as a case study. *J R Soc Interface*, 7(43):271–283, 2010.

[10] N. Bharti, A.J. Tatem, M.J. Ferrari, R.F. Grais, A. Djibo, and B.T. Grenfell. Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science*, 334(6061):1424–1427, 2011.

[11] C.E. Gunning and H.J. Wearing. Probabilistic measures of persistence and extinction in measles (meta)populations. *Ecol. Lett.*, 16:985–994, 2013.

[12] E. Sydenstricker and A.W. Hedrich. Completeness of Reporting of Measles, Whooping Cough, and Chicken Pox at Different Ages: Hagerstown Morbidity Studies: Supplement to Study No. II. *Public Health Rep*, 44(26):1537–1543, 1929.

[13] A.W. Hedrich. The corrected average attack rate from measles among city children. *Am. J. Epidemiol.*, 11(3):576, 1930.

[14] W.P. London and J.A. Yorke. Recurrent outbreaks of measles, chickenpox and mumps: I. Seasonal variation in contact rates. *Am. J. Epidemiol.*, 98(6):453, 1973.

[15] J.A. Clarkson and P.E.M. Fine. The efficiency of measles and pertussis notification in England and Wales. *Int J Epidemiol*, 14(1):153–168, 1985.

[16] E.A. Undurraga, Y.A. Halasa, and D.S. Shepard. Use of Expansion Factors to Estimate the Burden of Dengue in Southeast Asia: A Systematic Analysis. *PLoS Negl Trop Dis*, 7(2):e2056, 2013.

[17] F.L. Black. The role of herd immunity in control of measles. *Yale J Biol Med*, 55(3-4):351, 1982.

[18] O.N. Bjørnstad, B.F. Finkenstädt, and B.T. Grenfell. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecol Monogr*, 72(2):169–184, 2002.

[19] B. Grenfell and J. Harwood. (Meta) population dynamics of infectious diseases. *Trends Ecol. Evol. (Amst.)*, 12(10):395–399, 1997.

[20] B. Finkenstädt and B.T. Grenfell. Empirical determinants of measles metapopulation dynamics in England and Wales. *Proc. R. Soc. B*, 265(1392):211–220, 1998.

[21] N. Bharti, Y. Xia, O.N. Bjornstad, and B.T. Grenfell. Measles on the Edge: Coastal Heterogeneities and Infection Dynamics. *PloS ONE*, 3(4):e1941, 2008.

[22] A.J.K. Conlan, P. Rohani, A.L. Lloyd, M. Keeling, and B.T. Grenfell. Resolving the impact of waiting time distributions on the persistence of measles. *J R Soc Interface*, 7(45):623, 2010.

[23] C.F. Westoff. Differential Fertility in the United States: 1900 to 1952. *Am Sociol Rev*, 19(5):549–561, 1954.

[24] I. Nåsell. A new look at the critical community size for childhood infections. *Theor Popul Biol*, 67(3):203–216, 2005.

[25] P. Rohani, D. J. D. Earn, and B. T. Grenfell. Impact of immunisation on pertussis transmission in England and Wales. *Lancet*, 355(9200):285–286, 2000.

[26] F. S. Crum. A Statistical Study of Measles. *Am J Public Health*, 4(4):289–309, 1914.

[27] S.E. Tolnay. The African American Great Migration and Beyond. *Annu Rev Sociol*, pages 209–232, 2003.

[28] R.M. Anderson and R.M. May. *Infectious diseases of humans*. Oxford University Press Oxford, 1991.

[29] B.T. Grenfell and B.M. Bolker. Cities and villages: infection hierarchies in a measles metapopulation. *Ecol. Lett.*, 1(1):63–70, 1998.

[30] B.T. Grenfell, O.N. Bjørnstad, and B.F. Finkenstädt. Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. *Ecol Monogr*, 72(2):185–202, 2002.

[31] H.J. Wearing and P. Rohani. Estimating the duration of pertussis immunity using epidemiological signatures. *PLoS Pathog*, 5(10):e1000647, 2009.

[32] Central Intelligence Agency. The World Factbook, 2014.

[33] U.S. Public Health Service. Public Health Rep, 1920–1950. URL `http://www.ncbi.nlm.nih.gov/pmc/issues/149156/`.

[34] Office of Population Censuses and Surveys. Registrar General's weekly reports, England and Wales, 1948-1968.

[35] B. Bolker and B. Grenfell. Space, persistence and dynamics of measles epidemics. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 348(1325):309–320, 1995.

[36] P. Rohani, D.J.D. Earn, and B.T. Grenfell. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286(5441):968, 1999.

[37] U.S. Census Bureau. U.S. Census, 1920–1950.

[38] Linder, F.E. and Grove, R.D. Vital statistics rates in the United States, 1900-1940, 1947. URL `http://www.cdc.gov/nchs/products/vsus.htm`. Accessed June 2010.

[39] Grove, R.D. and Hetzel, A.M. Vital statistics rates in the United States, 1940-1960, 1968. URL `http://www.cdc.gov/nchs/products/vsus.htm`. Accessed June 2010.

[40] U.S. Bureau of the Census. Historical Statistics of the United States, Colonial Times to 1970, Bicentennial Edition, 1975. URL `https://www.census.gov/prod/www/statistical_abstract.html`. Accessed June 2010.

[41] Ruggles, S. and Alexander, J.T. and Genadek, K. and Goeken, R. and Schroeder, M.B. and Sobek, M. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database], 2010.

[42] P. Rohani. Personal communication, 2012.

[43] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1993.

Supplemental Figures and Tables

Incomplete reporting of pre-vaccine era childhood diseases: a case study of observation process variability

C. E. Gunning, E. Erhardt, and H. J. Wearing

Disease	Area	Era	N	Start	End
Measles	U.S. Cities	Early	574	1924-01-05	1934-12-29
Measles	E&W Cities	Early	286	1944-01-09	1954-12-26
Measles	U.S. Cities	Late	574	1935-01-05	1945-12-29
Measles	E&W Cities	Late	312	1955-01-09	1966-12-25
WC	U.S. Cities	Early	574	1924-01-05	1934-12-29
WC	U.S. Cities	Late	574	1935-01-05	1945-12-29

Table 3.3. Time range and sample number of subdivided city case reports. WC: whooping cough; E&W: England & Wales; N: number of sampled case report.

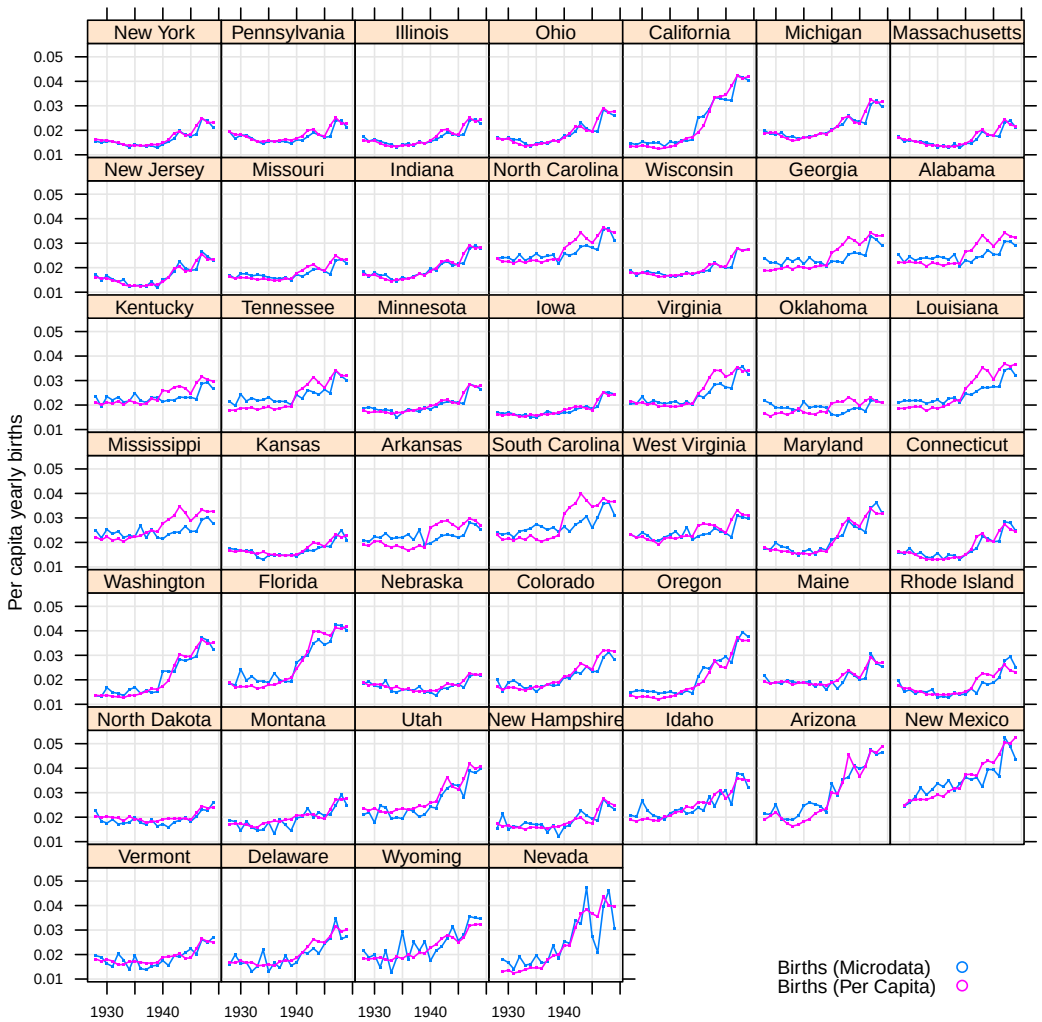

Figure 3.5. Comparison of per capita and microdata method by year. States are ordered by population size (1930, microdata method). The largest states show a close concordance between methods. The microdata method shows significant stochastic variation in smaller states.

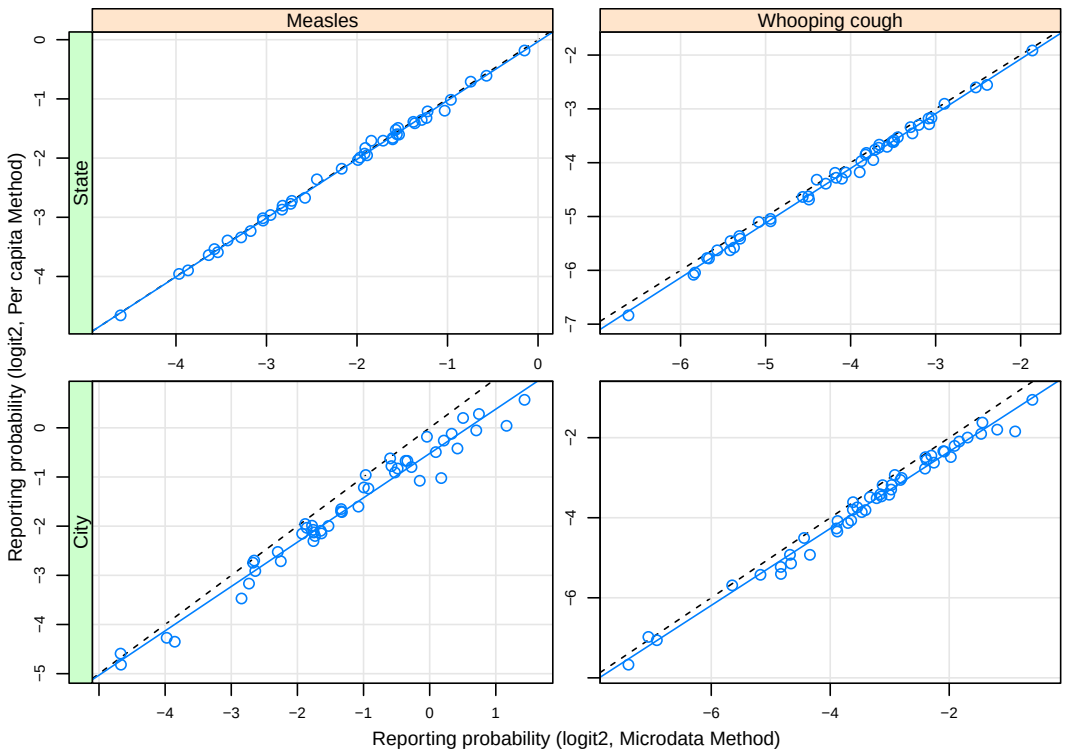

Figure 3.6. Comparison of estimated reporting probabilities between the per capita and the microdata method. For cities, the per capita method infers many more susceptibles, and thus lower reporting probabilities, than the microdata method. This is likely due to the use of state per capita birth rates to estimate city births. State birth rates include rural areas, which generally have higher per capita birth rates than urban areas (Figure 3.7). All values were logit_2 transformed to correct for heteroskedasticity. Black dashed line: 1-1 line; blue line: linear model.

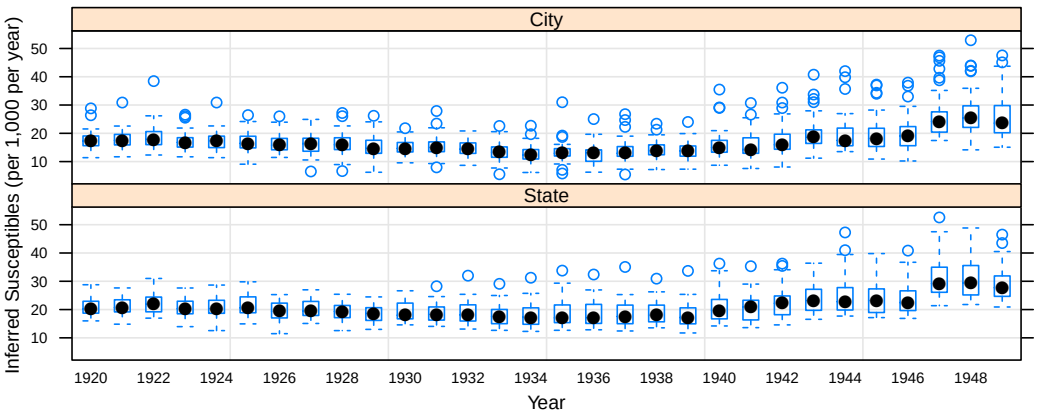

Figure 3.7. Distribution of yearly new susceptibles. The per capita rate of births + migration was inferred from IPUMS census records and the 1930 population size. Lower depression-era birth rates in the early 1930s and high post-war birth rates in the mid to late 1940s are apparent. Median yearly state birth rates are generally higher than associated city medians, likely due to the higher birth rates of rural populations.

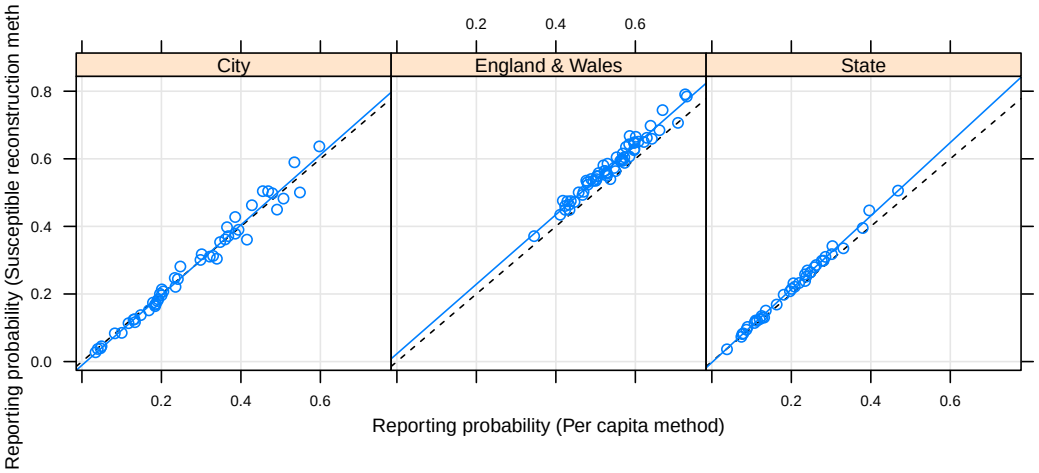

Figure 3.8. Comparison of reporting probability estimation methods. A close correspondence between the per capita method and the susceptible reconstruction method is evident (dashed black line: 1-1 line; blue line: linear model between the two methods).

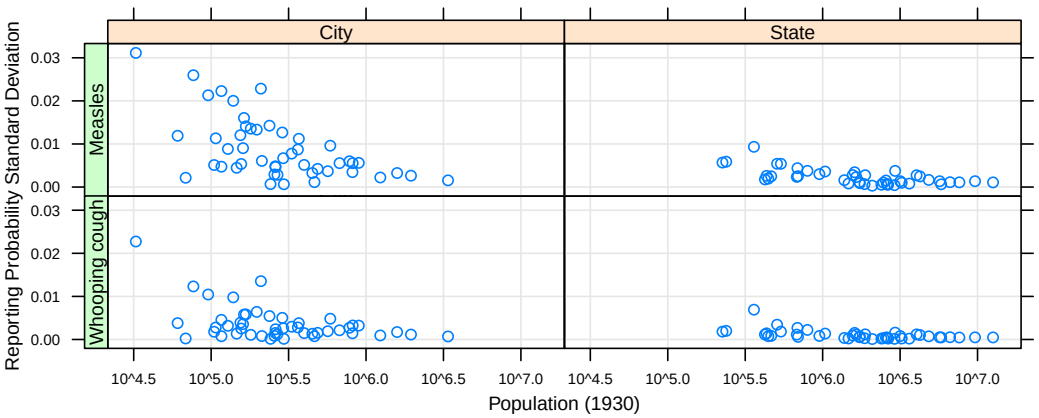

Figure 3.9. Standard error of reporting probability bootstrap draws, showing that variation decreases as the population of the sampled location increases. The 1930 population is shown, using the microdata method.

	Pop. (1930)	White (%)	Measles (%)	CI	W. cough (%)	CI
Alabama	2.6e+06	65	7.7	(7.6, 7.8)	2.1	(2.1, 2.1)
Arizona	4.3e+05	86	15.2	(14.7, 15.7)	6.2	(5.9, 6.4)
Arkansas	1.9e+06	74	7.2	(7.0, 7.3)	2.2	(2.1, 2.2)
California	5.7e+06	95	25.0	(24.7, 25.3)	7.0	(6.9, 7.1)
Colorado	1e+06	99	27.2	(26.6, 28.0)	7.5	(7.2, 7.8)
Connecticut	1.6e+06	98	28.9	(28.2, 29.6)	9.1	(8.8, 9.4)
Delaware	2.4e+05	85	20.5	(19.4, 21.7)	5.0	(4.7, 5.4)
Florida	1.5e+06	71	8.3	(8.1, 8.4)	1.8	(1.7, 1.8)
Georgia	2.9e+06	63	6.2	(6.1, 6.3)	1.6	(1.5, 1.6)
Idaho	4.5e+05	99	10.5	(10.1, 10.8)	2.9	(2.8, 3.1)
Illinois	7.6e+06	96	19.7	(19.5, 20.0)	5.9	(5.8, 6.0)
Indiana	3.2e+06	96	12.7	(12.5, 12.9)	2.1	(2.1, 2.2)
Iowa	2.5e+06	99	11.0	(10.8, 11.2)	2.3	(2.2, 2.3)
Kansas	1.9e+06	96	24.4	(23.9, 25.0)	7.5	(7.3, 7.7)
Kentucky	2.6e+06	91	8.9	(8.8, 9.0)	4.0	(3.9, 4.1)
Louisiana	2.1e+06	64	3.8	(3.8, 3.9)	0.9	(0.9, 1.0)
Maine	8e+05	100	24.7	(23.9, 25.4)	10.0	(9.6, 10.5)
Maryland	1.6e+06	83	22.6	(22.1, 23.0)	8.0	(7.8, 8.2)
Massachusetts	4.2e+06	99	33.0	(32.5, 33.5)	9.9	(9.7, 10.1)
Michigan	4.8e+06	96	26.8	(26.5, 27.1)	8.6	(8.5, 8.8)
Minnesota	2.6e+06	99	17.5	(17.2, 17.8)	3.8	(3.7, 3.9)
Missouri	3.6e+06	94	10.4	(10.3, 10.6)	1.8	(1.8, 1.9)
Montana	5.4e+05	97	27.9	(26.9, 29.0)	6.8	(6.4, 7.2)
Nebraska	1.4e+06	98	12.6	(12.3, 12.9)	1.9	(1.8, 2.0)
New Hampshire	4.7e+05	100	11.8	(11.3, 12.3)	2.7	(2.6, 2.9)
New Jersey	4e+06	95	35.6	(35.1, 36.2)	11.1	(10.9, 11.3)
New Mexico	4.2e+05	92	11.0	(10.6, 11.3)	5.3	(5.1, 5.6)
New York	1.3e+07	97	24.7	(24.5, 24.9)	7.6	(7.6, 7.7)
North Carolina	3.2e+06	72	20.4	(20.1, 20.6)	8.7	(8.5, 8.8)
North Dakota	6.8e+05	98	13.7	(13.3, 14.2)	4.6	(4.4, 4.8)
Ohio	6.6e+06	95	20.0	(19.8, 20.2)	6.7	(6.6, 6.8)
Oklahoma	2.4e+06	89	5.9	(5.8, 6.0)	1.7	(1.6, 1.7)
Oregon	9.5e+05	98	21.6	(21.0, 22.2)	4.3	(4.2, 4.5)
Pennsylvania	9.6e+06	95	28.7	(28.4, 28.9)	7.1	(7.0, 7.2)
Rhode Island	6.9e+05	98	23.7	(22.9, 24.6)	10.0	(9.5, 10.5)
South Carolina	1.7e+06	54	10.1	(10.0, 10.3)	5.8	(5.6, 5.9)
South Dakota	6.9e+05	97	14.2	(13.7, 14.7)	2.4	(2.3, 2.5)
Tennessee	2.6e+06	81	7.5	(7.4, 7.7)	2.9	(2.9, 3.0)
Texas	5.8e+06	86	12.1	(12.0, 12.2)	6.5	(6.4, 6.6)
Utah	5.1e+05	98	31.6	(30.6, 32.7)	14.7	(14.1, 15.4)
Vermont	3.6e+05	100	39.2	(37.4, 41.1)	20.3	(19.0, 21.7)
Washington	1.6e+06	97	25.1	(24.6, 25.6)	6.3	(6.2, 6.5)
West Virginia	1.7e+06	94	12.0	(11.8, 12.2)	4.0	(3.9, 4.1)
Wisconsin	2.9e+06	99	45.8	(45.1, 46.6)	13.7	(13.4, 14.1)
Wyoming	2.3e+05	97	20.7	(19.6, 21.8)	4.9	(4.6, 5.3)

Table 3.4. States: Demographic covariates and estimated reporting probabilities of each disease (showing median and 95% CI).

	Pop. (1930)	White (%)	Measles (%)	CI	W. cough (%)	CI
Atlanta, GA	2.7e+05	62	10.2	(9.6, 10.7)	4.2	(3.9, 4.4)
Baltimore, MD	8.2e+05	81	34.7	(33.7, 35.8)	19.6	(18.9, 20.2)
Birmingham, AL	2.6e+05	61	11.0	(10.5, 11.7)	3.0	(2.9, 3.2)
Boston, MA	7.8e+05	98	34.4	(33.2, 35.6)	15.3	(14.7, 15.8)
Bridgeport, CT	1.5e+05	98	11.1	(10.3, 12.0)	3.2	(3.0, 3.5)
Buffalo, NY	5.7e+05	98	18.4	(17.7, 19.1)	9.3	(9.0, 9.7)
Charleston, WV	6.1e+04	92	20.7	(18.6, 23.3)	6.4	(5.7, 7.2)
Chicago, IL	3.4e+06	93	18.2	(17.9, 18.5)	8.1	(7.9, 8.2)
Cincinnati, OH	4.5e+05	91	14.2	(13.6, 14.9)	5.7	(5.4, 5.9)
Cleveland, OH	9e+05	92	34.1	(33.1, 35.3)	18.7	(18.1, 19.4)
Columbus, OH	2.9e+05	86	23.5	(22.3, 24.9)	8.7	(8.2, 9.2)
Dallas, TX	2.6e+05	85	17.8	(16.9, 18.8)	5.4	(5.1, 5.7)
Denver, CO	2.9e+05	98	46.6	(44.2, 49.2)	17.8	(16.9, 18.8)
Detroit, MI	1.6e+06	92	28.3	(27.7, 29.0)	14.5	(14.2, 14.9)
Flint, MI	1.5e+05	96	33.7	(31.5, 36.2)	10.6	(9.9, 11.4)
Fort Wayne, IN	1.2e+05	96	11.4	(10.6, 12.4)	1.7	(1.6, 1.9)
Grand Rapids, MI	1.7e+05	99	38.0	(35.4, 40.9)	15.4	(14.3, 16.6)
Hartford, CT	1.6e+05	97	23.3	(21.6, 25.1)	8.8	(8.2, 9.6)
Houston, TX	2.9e+05	78	3.0	(2.8, 3.1)	0.7	(0.7, 0.7)
Indianapolis, IN	3.6e+05	87	36.1	(34.5, 37.9)	11.2	(10.6, 11.7)
Kansas City, MO	4e+05	91	20.1	(19.2, 21.2)	5.6	(5.4, 5.9)
Los Angeles, CA	1.2e+06	94	16.5	(16.1, 16.9)	6.8	(6.6, 6.9)
Memphis, TN	2.6e+05	60	17.9	(17.0, 18.9)	8.7	(8.3, 9.2)
Milwaukee, WI	5.9e+05	99	49.9	(48.1, 51.9)	24.1	(23.2, 25.1)
Mobile, AL	6.8e+04	67	4.7	(4.3, 5.2)	0.5	(0.5, 0.6)
Nashville, TN	1.6e+05	72	14.4	(13.4, 15.6)	6.6	(6.1, 7.1)
New Haven, CT	1.6e+05	95	41.9	(38.9, 45.2)	14.4	(13.3, 15.6)
New Orleans, LA	4.6e+05	71	5.4	(5.2, 5.6)	3.4	(3.3, 3.5)
Omaha, NE	2.1e+05	95	19.1	(17.9, 20.3)	2.4	(2.2, 2.6)
Philadelphia, PA	2e+06	88	24.0	(23.5, 24.6)	10.1	(9.9, 10.3)
Pittsburgh, PA	6.7e+05	92	29.8	(28.7, 30.9)	11.3	(10.9, 11.7)
Providence, RI	2.4e+05	98	46.0	(43.3, 48.9)	16.9	(15.8, 18.0)
Raleigh, NC	3.3e+04	61	42.6	(37.1, 49.2)	30.5	(26.5, 35.4)
Richmond, VA	1.8e+05	71	39.6	(37.1, 42.4)	3.0	(2.8, 3.2)
Rochester, NY	3.3e+05	100	27.0	(25.5, 28.6)	9.8	(9.3, 10.5)
Sacramento, CA	9.6e+04	93	43.3	(39.4, 47.8)	20.3	(18.4, 22.5)
Saint Louis, MO	8.2e+05	89	19.5	(18.8, 20.2)	7.7	(7.4, 8.0)
Salt Lake City, UT	1.4e+05	98	56.9	(53.3, 61.1)	26.8	(25.0, 28.9)
San Antonio, TX	2.4e+05	93	3.0	(2.8, 3.1)	0.6	(0.6, 0.7)
Seattle, WA	3.7e+05	96	41.8	(39.7, 44.1)	13.8	(13.0, 14.5)
South Bend, IN	1e+05	96	11.3	(10.4, 12.4)	3.9	(3.6, 4.3)
Spokane, WA	1.2e+05	99	50.6	(46.5, 55.3)	9.8	(9.0, 10.8)
Syracuse, NY	2.1e+05	100	63.0	(58.6, 67.6)	35.8	(33.4, 38.7)
Tacoma, WA	1.1e+05	97	26.6	(24.6, 29.0)	6.4	(5.9, 7.0)
Trenton, NJ	1.3e+05	94	19.2	(17.6, 21.0)	6.7	(6.2, 7.4)
Washington, DC	4.9e+05	72	20.8	(20.0, 21.6)	7.2	(6.9, 7.5)
Winston-Salem, NC	7.7e+04	56	53.3	(48.6, 58.8)	24.4	(22.1, 27.0)
Worcester, MA	2e+05	100	37.7	(35.3, 40.5)	17.4	(16.2, 18.7)

Table 3.5. Cities: Demographic covariates and estimated reporting probabilities of each disease (showing median and 95% CI).

Area	Response	DF	R^2	Parameter	Est.	Std. Error	t value	Pr($>|t|$)
State	Measles	41	0.515	(Intercept)	-2.25	0.105	-21.5	6.01e-24
				prop.white	5.84	1.06	5.5	2.23e-06
				prop.labforce	16.2	4.32	3.76	0.000538
				sd.housesize	2.33	0.849	2.75	0.0089
State	Whooping cough	41	0.404	(Intercept)	-4.18	0.12	-34.8	4.82e-32
				prop.white	6.63	1.33	4.99	1.14e-05
				sd.housesize	4.3	0.985	4.36	8.48e-05
				prop.labforce	13.6	5.01	2.72	0.00941
				prop.male	-18.7	9.74	-1.93	0.0612
City	Measles	45	0.317	(Intercept)	-1.25	0.17	-7.34	3.18e-09
				prop.school	36.8	8.04	4.58	3.68e-05
				mean.housesize	-1.66	0.646	-2.57	0.0135
City	Whooping cough	46	0.128	(Intercept)	-3.31	0.203	-16.3	9.31e-21
				prop.school	26.4	9.38	2.81	0.00721

Table 3.6. Linear models of reporting probability (microdata method) in response to demographic covariates. Models were constructed via forward selection with a BIC selection criteria. For each model, parameters are shown in decreasing order of BIC reduction. Separate models are denoted by horizontal lines. Regardless of disease, proportion white is the best predictor of state reporting probabilities, while proportion attending school is the best predictor of city reporting probabilities. Reporting probability was logit_2 transformed and all predictors were zero-centered prior to model construction.

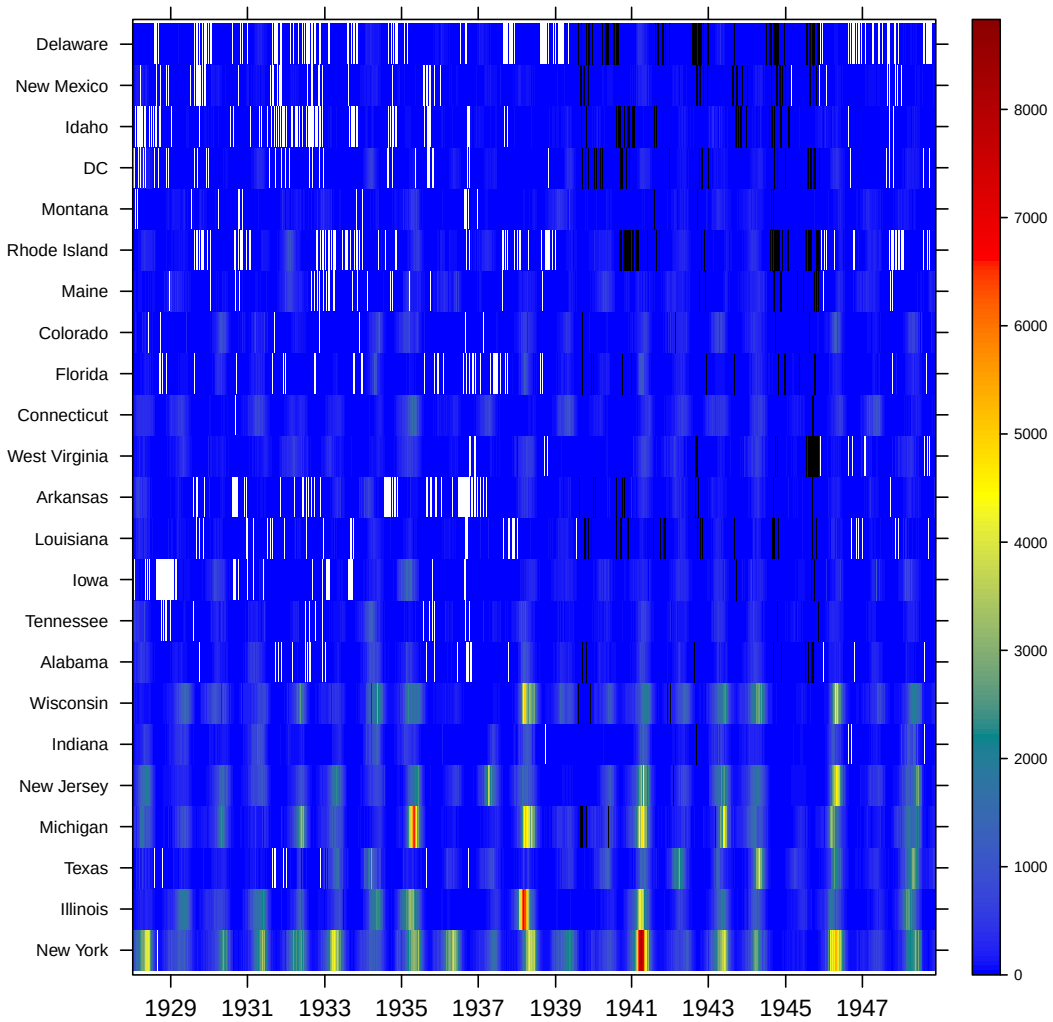

Figure 3.10. Measles, State. Case reports per sample period, with locations ordered by population size (black = 0; white = missing). Every other location is shown.

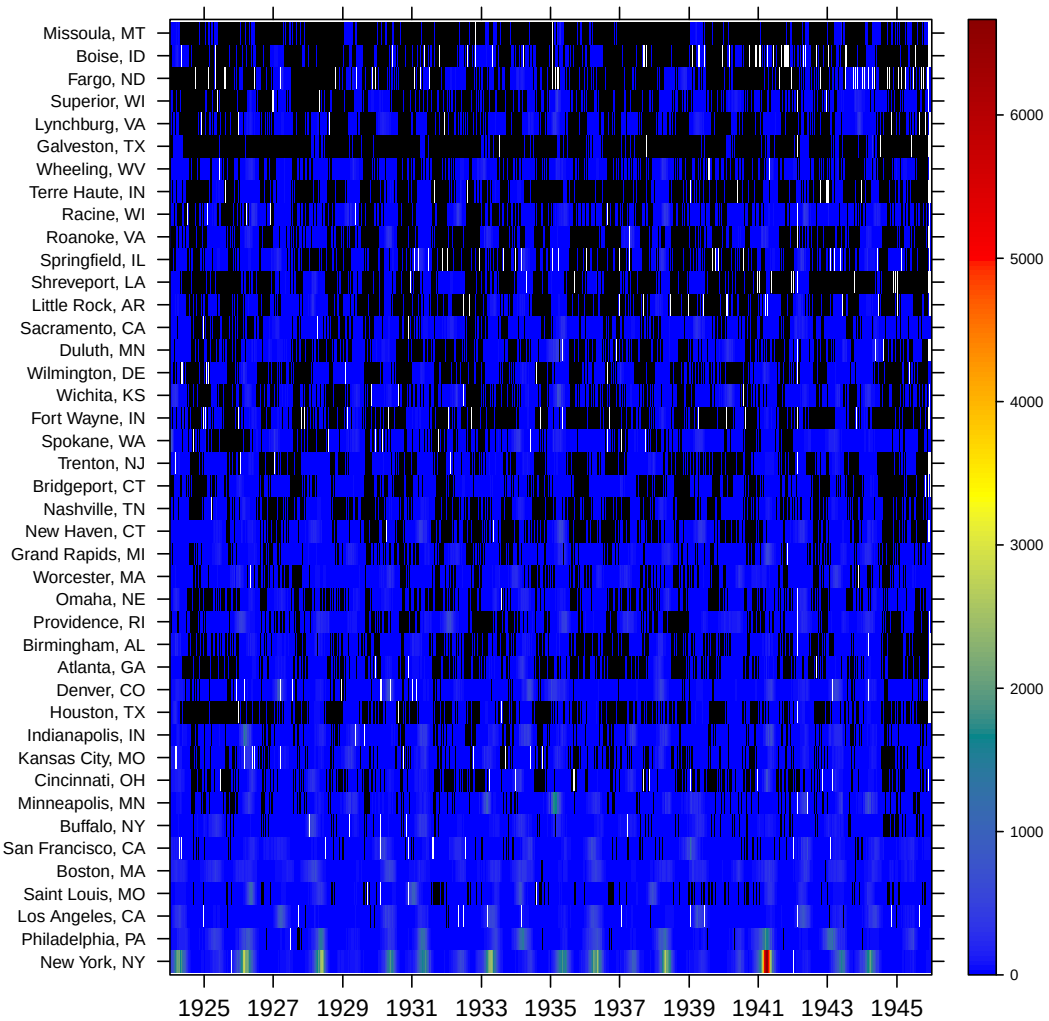

Figure 3.11. Measles, City. Case reports per sample period, with locations ordered by population size (black = 0; white = missing). Every other location is shown.

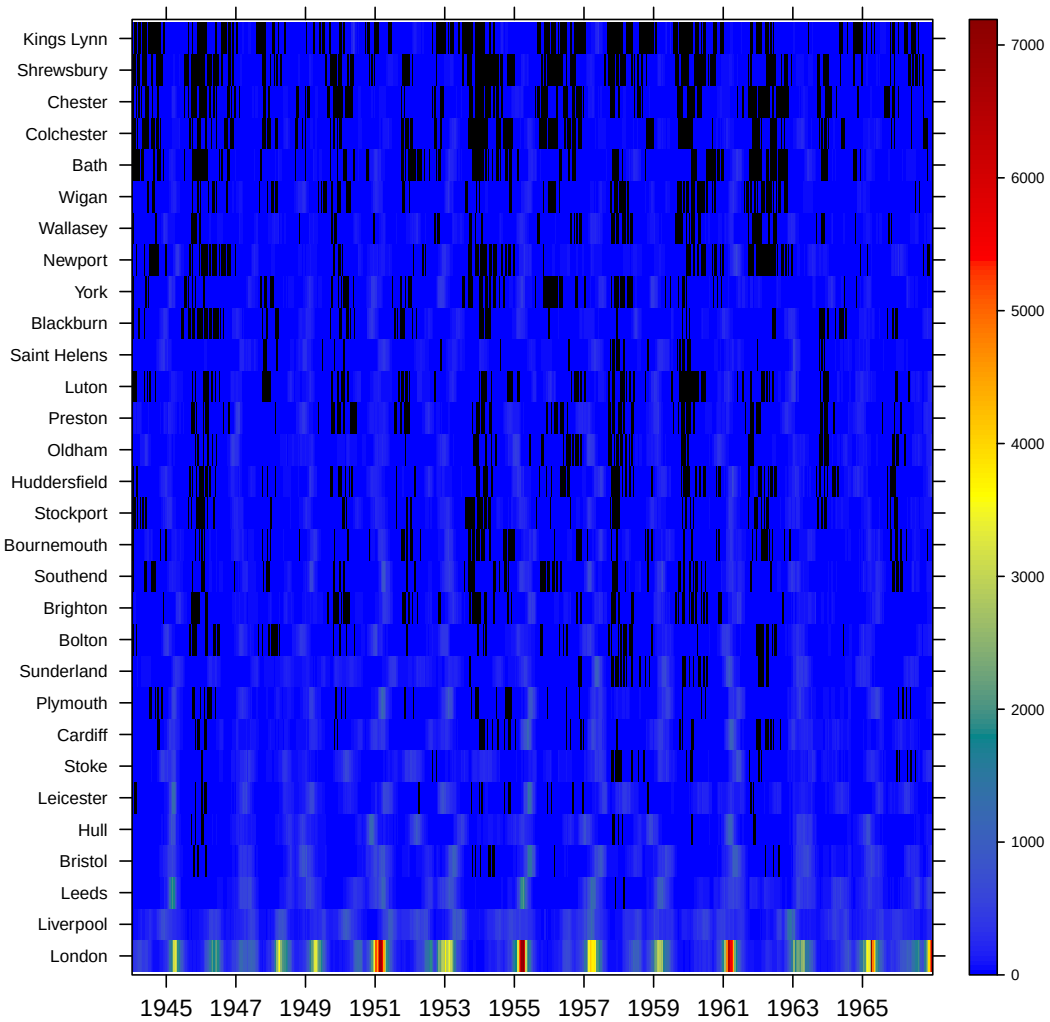

Figure 3.12. Measles, England & Wales. Case reports per sample period, with locations ordered by population size (black = 0; white = missing). Every other location is shown.

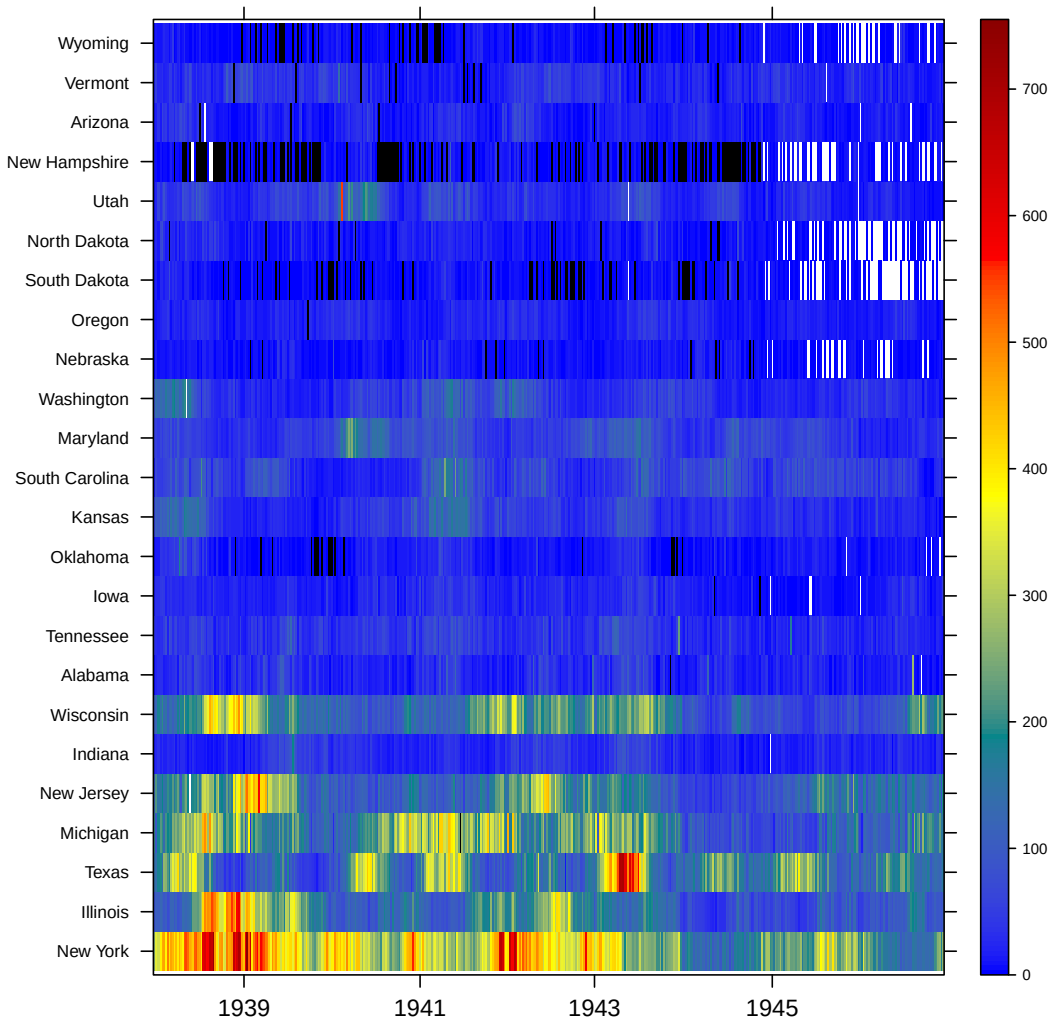

Figure 3.13. Whooping cough, State. Case reports per sample period, with locations ordered by population size (black = 0; white = missing). Every other location is shown.

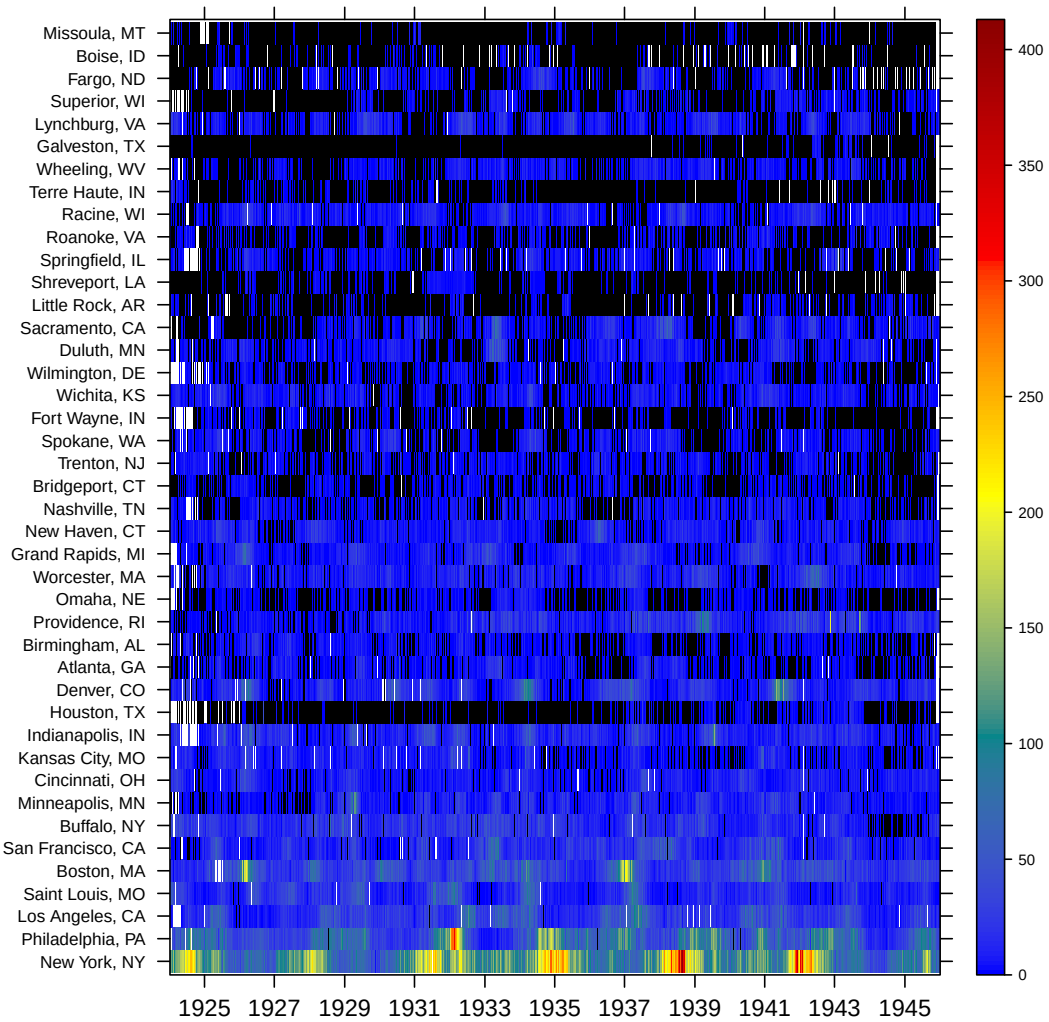

Figure 3.14. Whooping cough, City. Case reports per sample period, with locations ordered by population size (black = 0; white = missing). Every other location is shown.

Chapter 4

Predicted and Cryptic Persistence

Cryptic persistence in childhood disease

Christian Gunning[1,*] Matthew Ferrari[2], Helen J. Wearing[1,3]

1 Department of Biology, University of New Mexico, Albuquerque, New Mexico, USA

2 Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, Pennsylvania, USA

3 Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA

∗ **E-mail: xian@unm.edu**

4.1 Abstract

Accurate monitoring of disease incidence is a key element in the control of preventable infectious diseases. Previous work has demonstrated wide variation in disease reporting probability. Here we use an extensive record of measles and whooping cough case reports in pre-vaccine era U.S. cities to predict disease persistence from report-

ing probability and distributions of case reports. Our results indicate that cryptic persistence is common in populations with a combination of imperfect reporting and low absolute incidence. In turn, both population size and pathogen life history affect patterns of absolute incidence, independent of disease reporting. Thus, we find that cryptic persistence is most common in medium-sized cities for measles, and smaller cities for whooping cough. As modern vaccination campaigns push absolute incidence to lower levels, cryptic persistence is expected to increase, particularly in "colonizer" diseases with longer infectious periods and lower transmission rates.

4.2 Introduction

Persistence of childhood diseases such as measles and whooping cough results from a complex interplay between population and metapopulation processes. At the local level, stochasticity in host and pathogen demographic processes can result in local extinction, particularly in small populations [1–4] or for pathogens with short infectious periods [4, 5]. At the metapopulation level, connectivity between populations allows the rescue of individual populations from local extinction [5, 6], while the effect of connectivity on temporal synchrony and metapopulation persistence is less clear [7–12]. Disentangling local and metapopulation processes has proved challenging. Incomplete observation further complicates the picture: measles has unambiguous symptoms that remain approximately constant with age, and generally has higher and less variable reporting probabilities than whooping cough [13], which exhibits age-dependent severity and shares symptoms with many other common respiratory diseases [14, 15].

Here we compare patterns of persistence between measles and whooping cough in pre-vaccine era United States cities. The pre-vaccine U.S. provides an attractive model system with high-quality demographic records, two decades of continuous

disease monitoring in the majority of urban areas, and no uncertainty associated with vaccine uptake and efficacy. We are also able to correct for variation in reporting probability.

Local Drivers of Persistence

Measles and whooping cough are diseases caused by acutely infectious and fully immunizing obligate human pathogens. The pathogens that cause these so called "childhood diseases" infected an overwhelming majority of the human population in the pre-vaccine era, with infection typically occurring early in life. Periodic forcing of disease transmission via changes in host density (via, e.g., school terms [16, 17] or economic migration [18]) is also a common feature of childhood diseases. Both measles and pertussis have high reproductive ratios ($R_0 \approx 20$) [19] and relatively fast life cycles: the combined latent and infectious period is approximately 15 days for measles and 30 days for pertussis [17]. At high incidence, susceptible hosts are rapidly depleted, leading to subsequent inter-epidemic troughs of low incidence or stochastic extinction. When infection is low or absent from a population, susceptible replenishment proceeds via the host demographic processes of birth and migration. These forces combine with stochasticity to yield characteristic yearly and multi-annual epidemic cycles in a wide range of human populations and diseases [9, 20–26].

During inter-epidemic troughs, stochastic extinction of measles and pertussis is common, particularly in small populations. Indeed, previous work has shown that persistence scales approximately log-linearly with population size [1, 4, 27–31]. Theory predicts that longer infectious periods and higher birth rates should increase persistence when all else is equal [4, 27].

Vaccination programs remove susceptible individuals from the chain of infection, and are analogous to reducing birth rates [23]. Yet vaccination is imperfect in both

application and protective effects. In addition, the immunizing effects of natural infection decrease as incidence drops [32], which can lead to a paradoxical negative feedback between vaccine-induced immunity and immunity from natural infection.

Metapopulation Drivers of Persistence

Persistence is typically measured in individual populations, yet it both influences and is influenced by metapopulation processes. Pathogen reintroduction via importation restarts (or "rescues") local chains of infection from extinction [5, 31, 33] (note that we define persistence as the presence of disease, either from local transmission or importation). Increased metapopulation connectivity can increase rescue effects, leading to increased persistence throughout a metapopulation [6]. On the other hand, low connectivity and prolonged extinction leads to susceptible build-up and (eventual) explosive epidemics.

The effect of metapopulation synchrony on persistence remains an area of active research. Theory predicts that metapopulation asynchrony of disease incidence should increase persistence by fostering disease importation during local epidemic troughs [7, 12, 34, 35]. Inversely, high levels of connectivity can synchronize populations, leading, counter-intuitively, to decreased rescue effects and decreased local persistence in inter-epidemic troughs of metapopulation incidence [11, 12, 34, 36, 37]. Indeed, vaccination programs can drive metapopulation synchronization [9], which is hypothesized to favor extinction [8, 11, 38].

Outline

Here we examine metapopulation patterns of persistence of two childhood diseases in U.S. cities in the pre-vaccine era using an extensive dataset of measles and whooping

cough case reports, demographic records, and estimates of reporting probability. This study system allows us to test the relative effects of disease life history on persistence, providing a comparison between a relatively "invasive" pathogen (measles) versus a superior "colonizer" (whooping cough).

We focus first on observed persistence: that is, the proportion of non-zero case reports in each city over the period of record. We demonstrate a conserved scaling relationship between observed incidence (case reports), disease reporting probability, and observed persistence. We use this relationship to predict true persistence (defined as the expected persistence at full reporting). We examine the dependence of observed and predicted persistence on population size, and provide empirical support for the increased persistence of whooping cough relative to measles across a wide range of population sizes.

We next focus on cryptic persistence, which we define as the difference between observed and predicted persistence: the predicted probability of unobserved persistence. We explore the scaling of cryptic persistence with both population size and reporting probability. We show that cryptic persistence is particularly common in populations with low absolute incidence and low reporting probability, and thus differs markedly between diseases.

Our results suggest that cryptic persistence is widespread, and that metapopulation persistence is much more difficult to measure than previously acknowledged. In addition, vaccination appears capable of either decreasing or increasing cryptic persistence, depending on population and disease attributes such as reporting probability and infectious period. The "unknown unknowns" intrinsic in cryptic persistence complicates public health and epidemiology decision-making, with metapopulation persistence likely becoming more difficult to estimate as control efforts improve. Consequently, active disease surveillance plays an important public health role, even in the face of apparent disease elimination.

4.3 Methods

Estimating the Distribution of Incidence

Our ultimate goal is to estimate the probability of true persistence from case reports and reporting probability. This, in turn, allows us to compare metapopulation patterns of true and cryptic persistence between diseases. To accomplish this, we first infer the distribution of incidence from case reports and reporting probabilities, and then use a statistical model to estimate persistence from incidence. Throughout, we distinguish between inferred quantities and approximated or predicted quantities. Inferred quantities depend on distributional assumptions and are denoted by the *hat* superscript: $\widehat{X}$. Approximations and predictions, on the other hand, have simple arithmetic relationships to the true, unobserved quantities that they estimate, and are denoted by the *tilde* superscript: $\widetilde{X}$.

One key challenge we face is how to summarize distributions of incidence in a way that can be applied to both large and small populations alike. At large population sizes and over long time periods, both cases (true incidence) and case reports (observed incidence) follow log-normal distributions in these diseases. Small and medium sized populations, however, exhibit both local extinction and unobserved incidence due to incomplete reporting. These zero-observations require careful consideration, since $log(0)$ is undefined. If zero-observations are removed, then the estimated distribution is simply the distribution of case reports conditional on both persistence and subsequent observation of persistence.

One potential distributional measure is the sample median, which is transformation-invariant, and converges to the sample mean for log-normally distributed data (i.e., in large populations). However, the effects of discretization are especially apparent in small populations. Numerous cities have zero median case reports but

widely differing distributions of true incidence and case reports, making the median a poor estimator of either true incidence or case reports in small and medium sized cities.

Consequently, we use a distributional approach to infer the mean of case reports using nonlinear minimization to fit a normal CDF to the ECDF of log case reports. Since case reports of zero can arise in multiple ways (i.e., either from local extinction or cryptic persistence), we treat zeros as less than or equal to 1 case report. This process is equivalent to the method employed by Gunning and Wearing [33], applied here to case reports (i.e., observed absolute incidence rather than predicted per capita incidence). Thus we find the inferred mean of case reports $\widehat{\mu_c}$, which summarizes the full distribution of case reports (i.e., observed incidence).

In the largest populations, all three measures converge, while the log-space sample mean exhibits a positive bias in most populations due to the exclusion of zero observations (Figure 4.4). The sample median generally agrees with $\widehat{\mu_c}$, but is discrete and achieves a minimum of 0 in more than 10 cities for each disease. Thus, as the distribution of observed incidence shifts leftwards (typically in smaller populations), the sample median provides a decreasingly *precise* summary of observed incidence compared to $\widehat{\mu_c}$.

Incomplete observation

Disease reporting in this metapopulation is known to vary with both disease identity and location, and is approximately stationary over time in this system [13]. For each population i and each disease j, a single reporting probability r_{ij} was estimated from the ratio of births and the sum of case reports over the period of observation. Births were estimated from U.S. census microdata, which also permits the construction of confidence intervals through bootstrapping [13].

The first step to estimating persistence is to find the marginal distribution of true incidence (I) over time for each population and disease. First, we infer the marginal distribution of observed incidence (i.e. case reports, C). We then correct for incomplete reporting to yield an inferred distribution of incidence, which is equivalent to the inferred distribution of case reports at full reporting.

The expected or predicted incidence $\widetilde{I}$ is simply the observed case reports divided by the reporting probability: $\widetilde{I} = \frac{C}{r}$. This correction fails, however, for $C = 0$. $Pr(I = 0|C = 0) \propto Pr(C = 0|I = 0) * Pr(I = 0) = 1 * Pr(I = 0)$. Yet we seek $Pr(I = 0)$. That is, we wish to know the *true* probability of extinction ($I = 0$). Given $C = 0$, the maximum likelihood estimator (MLE) of I is 0. However, for low reporting probability and low incidence, a large proportion of observed zeros result from cryptic persistence. Assuming a binomial reporting process, $Pr(C = 0|r = \rho, I = i) = (1 - \rho)^i$ and, for $r = 0.1$ and $I = 10$, more than 30% of observed zeros are expected to result from cryptic persistence. We sidestep this complication by computing the inferred mean of incidence from the inferred mean of case reports (which is non-zero by design): $\widehat{\mu_I} \approx \frac{\widehat{\mu_c}}{r}$. The result, $\widehat{\mu_I}$, yields a summary statistic of the distribution of case reports at full reporting, which in turn is equivalent to the distribution of true incidence.

Predicting Persistence

The probability of persistence depends on season, and potentially on metapopulation incidence. Here we marginalize over time and focus on long-term differences in persistence between diseases and populations. Thus, we estimate the quasi-stationary, per-time probability of true persistence. We define *true persistence* (P) as the per-observation probability of non-zero incidence, and observed persistence (P_o) as the probability of non-zero case reports (e.g. the proportion of non-zero reporting weeks).

We construct a statistical model where observed persistence varies in response to the inferred mean of case reports: $log_{10}(Arctanh(P_o)) \sim log_{10}(\widehat{\mu_c})$ (see Figure 4.5). We use this model to estimate true persistence ($\widetilde{P} \approx P$) from the inferred distribution of incidence (which is equivalent to the inferred distribution of case reports at full reporting): $\widetilde{P} \sim \widehat{\mu_I}$ (see Figure 4.6).

Due to incomplete reporting, observed persistence is the sum of true persistence and cryptic persistence (P_c): $P_o = P + P_c$. Thus, cryptic persistence is the (unknown) fraction of unobserved incidence. Our best estimate of cryptic persistence is, then, the difference between observed and predicted persistence: $P_o - \widetilde{P} = \widetilde{P_c}$.

Estimating uncertainty

Numerous sources of error and variation are present in the final estimates of persistence and cryptic persistence. The primary sources of uncertainty in persistence estimates include the reporting probability and uncertainty in linear model predictions. We used parametric bootstrapping to quantify this uncertainty.

For each bootstrap draw, reporting probability was parametrically sampled and used to compute $\widehat{\mu_I}$ from $\widehat{\mu_C}$. Next, $log_{10}(Arctanh(\widetilde{P}))$ was parametrically sampled from the prediction distribution of the appropriate linear model, conditioned on the sampled $\widehat{\mu_I}$, and back-transformed into a proportion. Finally, $\widetilde{P_c}$ was computed from P_o and the sampled $\widetilde{P}$. Bootstrap samples were then used to construct 95% confidence intervals for $\widetilde{P}$ and $\widetilde{P_c}$.

4.4 Results

Regardless of disease, we clearly expect to observe fewer zero-weeks as mean case reports increase (Figure 4.1: $\widehat{\mu_C}$ versus P_o and $\widetilde{P}$). Indeed, the scaling of observed persistence (P_o) with observed incidence ($\widehat{\mu_C}$) is very similar between diseases (see also Figure 4.5). Yet theory predicts that pertussis, with a longer infectious period and a lower transmission rate, should exhibit less frequent stochastic extinction than measles for a given population size [4, 27], a pattern that is obscured by whooping cough's low reporting probability. Correcting for incomplete reporting shows that whooping cough is indeed much more likely to persist ($\widetilde{P} = Pr(I > 0)$) across a wide range of observed incidence. In addition, whooping cough displays much greater differences between observed and predicted persistence due to lower reporting probabilities.

For both diseases, increased persistence is expected with increasing population size [1–4]. Variable reporting probabilities again obscure any such scaling (Figure 4.2A), while the expected scaling is clearly seen once variable reporting is corrected for (Figure 4.2B). As theory predicts, whooping cough also exhibits increased persistence compared to measles across a range of population sizes (Figure 4.2B). In measles, cryptic persistence peaks in medium-sized cities that teeter on the edge of extinction, and is less common in smaller cities that exhibit more frequent local extinction. In whooping cough, on the other hand, smaller populations rarely exhibit local extinction and commonly exhibit low absolute incidence, making cryptic persistence common (Figure 4.2C). At large populations, we find that cryptic persistence is rare in both diseases (Figure 4.2C).

We expect lower reporting probabilities to yield increases in cryptic persistence, which we observe in both diseases (Figure 4.3). For a given reporting probability, larger populations also generally exhibit less cryptic persistence than smaller cities,

regardless of disease. Again, cryptic persistence is essentially absent in the largest cities, regardless of disease or reporting probability (Figure 4.3). Yet we find marked differences between diseases. For a given reporting probability, measles generally experiences much higher cryptic persistence, likely due to lower absolute incidence.

4.5 Discussion

Despite widespread availability of inexpensive and effective vaccines, childhood diseases have resisted elimination efforts. Classic epidemiological theory proposes that reducing the susceptible proportion of a population below $\frac{1}{R_0}$ should interrupt disease transmission, leading to local extinction [19]. Yet metapopulation elimination of disease has proven elusive and expensive: morbidity and mortality from vaccine-preventable diseases remains high in developing nations [39, 40], and importation of infection back into previously disease-free populations and metapopulations continues [41–43].

Where, when, and why vaccine-preventable diseases persist remain key ecological questions with important modern epidemiological consequences. As we have shown, incomplete disease reporting substantially affects common measures of persistence, particularly for low reporting probability and low absolute incidence. This impedes inference about disease dynamics at the local scale, and complicates comparisons between diseases or metapopulations with different reporting probabilities.

Here we extend previous work that employed distribution-based inference of disease incidence, and accounted for incomplete reporting [13, 33], to estimate the true persistence of both measles and whooping cough. We show that incomplete and variable disease reporting in this metapopulation obscures large-scale patterns of disease persistence.

For example, a naive measure of critical community size (CCS) such as the log-linear scaling of observed persistence with population size (Figure 4.1A) cannot distinguish between measles and whooping cough, while accounting for incomplete reporting clearly reveals whooping cough's higher persistence (Figure 4.1B). Population thresholds of extinction such as CCS have been criticized as poorly specified and difficult to measure [4, 33, 44]. Nonetheless, CCS remains a commonly reported feature of empirical data. Using a simple empirical definition (CCS := $min(\text{Population})|P > 0.95$; see also Figure 4.7), we find the measles CCS changing from $\approx$ 600 thousand (P_o) to $\approx$ 300 thousand ($\widetilde{P}$), while the whooping cough CCS changes from $\approx$ 200 thousand (P_o) to $<$ 100 thousand ($\widetilde{P}$), again highlighting the sizable effects of incomplete reporting.

Despite large differences in reporting probabilities, we find that cryptic persistence is widespread in both diseases. We expect that cryptic persistence is concentrated in cities that exhibit long periods of low but non-zero incidence, teetering on the edge of stochastic extinction. Yet the characteristics of these "refuge" populations differ markedly between disease. We find that cryptic persistence is concentrated at medium-sized populations in measles, and at smaller populations in whooping cough (Figures 4.2C and 4.3C). This accords with epidemiological theory, which predicts that the measles' high transmission rate and short infectious period leads to rapid susceptible depletion in small populations. Thus, small populations are expected to more commonly exhibit *true* extinction of measles. Whooping cough, on the other hand, can sustain low but non-zero incidence in much smaller populations than measles due to a longer infectious period and lower transmission rate.

One key challenge in disease ecology is unraveling the complex feedbacks between persistence in a metapopulation and its individual populations. Local persistence is driven both by local processes (birth, disease transmission) and metapopulation processes (host migration, disease importation). Thus, lowering metapopulation

incidence should, in general, decrease local persistence by reducing disease importation. How local persistence scales up to metapopulation persistence is less clear. Conventional epidemiological wisdom [45, 46] holds that metapopulation persistence depends on local persistence in focal cities above a critical size (CCS). Recent work suggest that aggregates of medium-sized cities exhibit patterns of persistence similar to individual cities of comparable size [33]. Evidence of widespread cryptic persistence in cities that commonly exhibit low absolute incidence further emphasizes the role that "non-focal" cities can play in metapopulation persistence.

4.6 Broader Applications

The dependence of cryptic persistence on both reporting and absolute incidence has important implications for modern disease control efforts. Cryptic persistence is uncommon at high absolute incidence, regardless of reporting probability. Yet as control measures drive down incidence, cryptic persistence becomes more and more sensitive to incomplete reporting. The overall effect is that, at low to intermediate reporting probabilities, the disease state of a local population (persistent or extinct) becomes less and less certain as local disease elimination is approached. Further, this effect varies by disease life history, as longer infectious periods favor persistence over extinction. Whooping cough, with historically low reporting probabilities and a long infectious period relative to other epidemic diseases, appears especially prone to cryptic persistence in populations where vaccination is incomplete.

For example, as measles incidence drops due to increased vaccination, we expect to observe local extinction for measles at higher population sizes (i.e. cryptic persistence shifting upwards from mid-sized populations, Figure 4.2C). Failure to account for cryptic persistence could lead to biased assessment of the success of control efforts, and mistaken allocation of control efforts away from areas where measles still persists.

Thus, more active surveillance to verify local extinction in medium-sized populations might be warranted. By comparison, whooping cough is more likely to exhibit cryptic persistence in small populations, suggesting that enhanced surveillance in these areas might be necessary to verify local extinction.

The interaction between cryptic persistence and natural immune boosting is another possible "unknown unknown". The well-known "honeymoon period" [47] refers to combined benefits of disease-induced and vaccine-induced immunity in a population shortly after the introduction of vaccination. As disease incidence falls, however, disease-induced immunity drops and higher levels of vaccination are required to achieve disease control. Here we have estimated cryptic persistence rather than cryptic incidence. Nonetheless, the extent to which cryptic disease incidence induces natural immunity or immune boosting [15, 48] warrants further attention. Serological tests and vaccines that allow the differentiation of vaccine-derived and naturally induced immunity, like those currently used in animal systems [49, 50], would greatly aid in resolving these ambiguities.

Cryptic persistence has important modern implications for the evolution of drug-resistance in parasites such as *Mycobacterium tuberculosis* and *Plasmodium falciparum*, where unobserved incidence can provide an important reservoir for resistant strains. Tuberculosis presents a long-standing problem, where poor sensitivity of diagnostic tests, difficulty distinguishing between latent and active infections, and socioeconomic limitations to diagnosis and treatment are all well-recognized problems [51–55]. Recent work has revealed that previously-unidentified asymptomatic malaria infection occurs extensively on the China-Myanmar border, where artemisinin-resistant *Plasmodium falciparum* has been detected [56–59].

The role that socioeconomic instability can play in cryptic persistence also warrants the attention of epidemiologists and public health officials. Regional conflicts can destabilize public health systems, leading to a simultaneous decrease in control

measures and disease monitoring, as well as large-scale increases in economic migration. The current global push to eradicate polio provides an illustrative modern example of the role conflicts play in disease prevention and monitoring [60, 61]. In this case, polio persists in the politically volatile regions of Nigeria and Pakistan, while more recent outbreaks have occurred in war-torn Somalia and Syria [62].

We hope that the results presented here will encourage public health professionals and epidemiologists to *anticipate* and proactively account for cryptic persistence. As part of routine disease control efforts, or in a high-stakes disease eradication campaign, identifying the populations most at risk of cryptic persistence can aid in the effective allocation of limited resources.

4.7 Figures

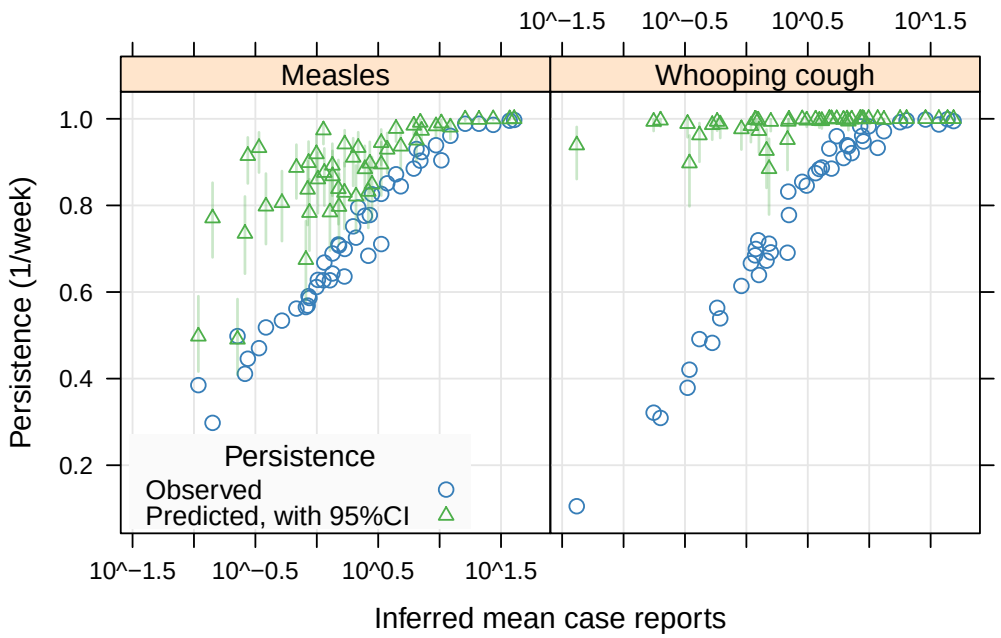

Figure 4.1. Observed (P_o, blue circles) and predicted ($\widetilde{P}$, green triangles) persistence versus inferred mean case reports ($\hat{\mu_C}$). We define observed persistence as the long-time, per-week probability of non-zero case reports; ($\hat{\mu_C}$) is a proxy for observed incidence. For each disease, a linear model (Figure 4.5) was used to predict persistence from incidence assuming 100% reporting. In general, the difference between observed and predicted persistence decreases with increasing incidence. 95% bootstrap confidence intervals of prediction are shown. See Figure 4.6 for details.

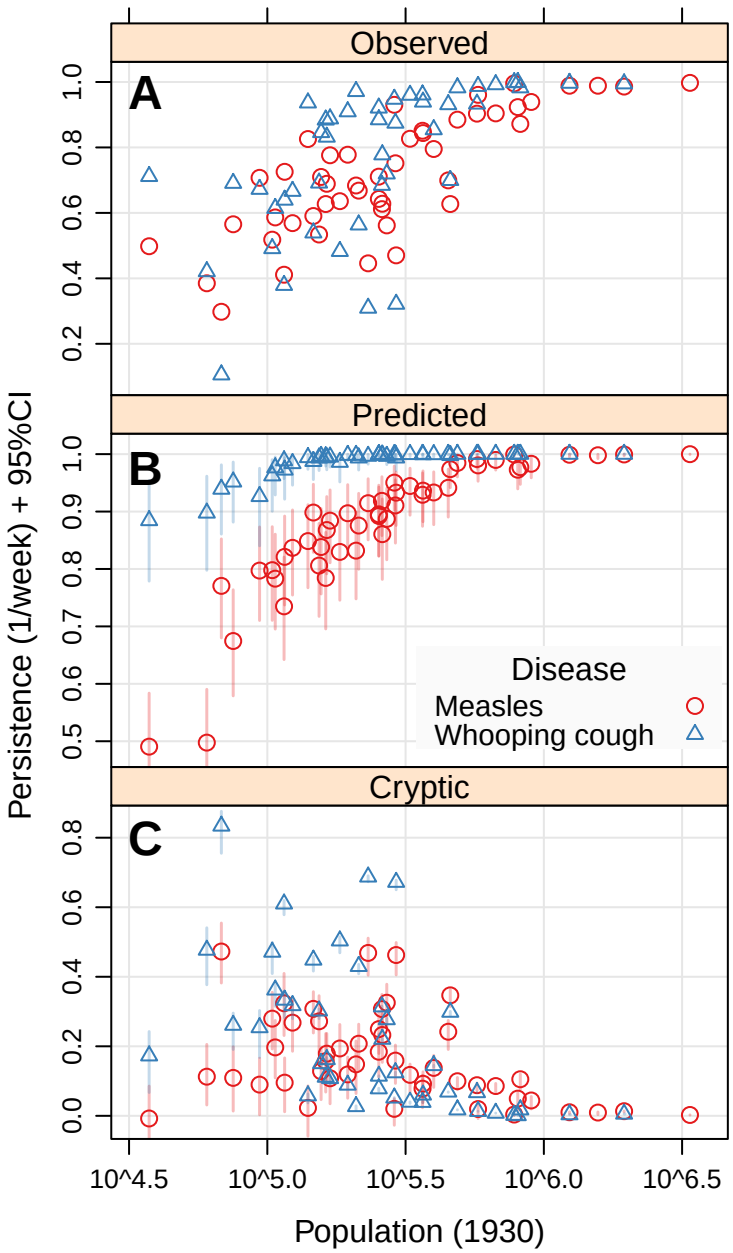

Figure 4.2. Observed, predicted, and cryptic persistence by population size, with 95% confidence intervals of predictions. Whooping cough shows greater predicted persistence than measles across a range of population sizes, though no difference between diseases is evident in observed persistence. Cryptic persistence is more common in small to medium sized cities due to low absolute incidence, and more common in whooping cough than in measles due to poor reporting.

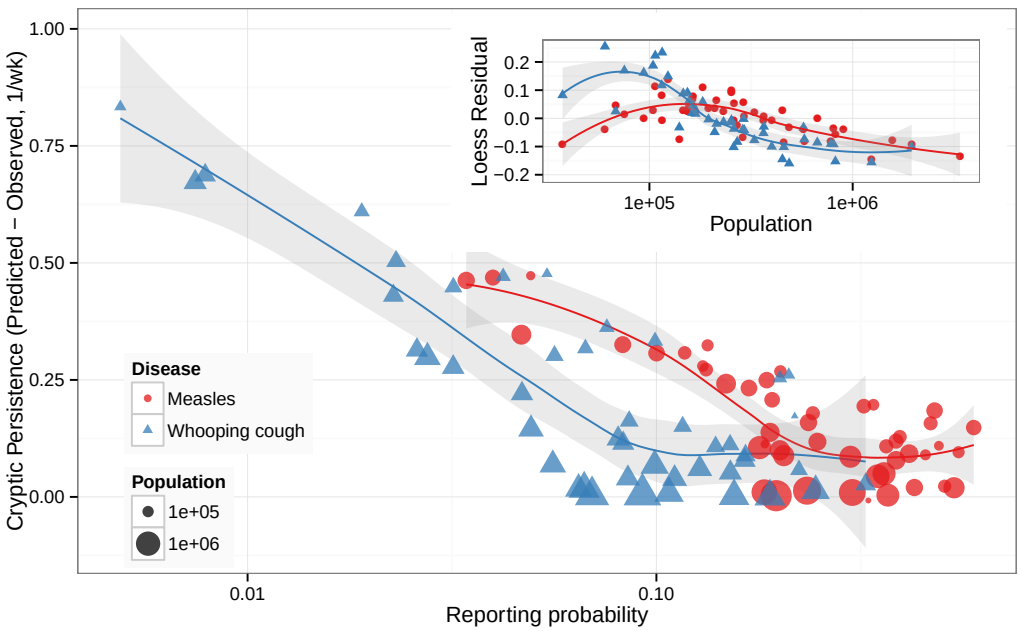

Figure 4.3. Cryptic persistence by reporting probability, showing more complete reporting of measles. A superimposed loess regression shows that, at low reporting probabilities, cryptic persistence is strongly correlated with reporting probability. The residuals of the loess regression are also shown (inset figure). For a given reporting probability, larger cities generally exhibit less cryptic persistence than smaller cities, particularly for whooping cough. Cryptic persistence is essentially absent in the largest cities, regardless of disease or reporting probability.

4.8 References

[1] M.S. Bartlett. The critical community size for measles in the United States. *J R Stat Soc Ser A*, 123(1):37–44, 1960.

[2] F.L. Black. Measles endemicity in insular populations: critical community size and its evolutionary implication. *J. Theor. Biol.*, 11(2):207–211, 1966.

[3] M.J. Keeling and B.T. Grenfell. Disease extinction and community size: modeling the persistence of measles. *Science*, 275(5296):65, 1997.

[4] I. Nåsell. A new look at the critical community size for childhood infections. *Theor Popul Biol*, 67(3):203–216, 2005.

[5] C.J.E. Metcalf, K. Hampson, A.J. Tatem, B.T. Grenfell, and O.N. Bjørnstad. Persistence in Epidemic Metapopulations: Quantifying the Rescue Effects for Measles, Mumps, Rubella and Whooping Cough. *PloS ONE*, 8(9):e74696, 2013.

[6] I. Hanski. Metapopulation dynamics. *Nature*, 396(6706):41–49, 1998.

[7] A.L. Lloyd and R.M. May. Spatial Heterogeneity in Epidemic Models. *J. Theor. Biol.*, 179(1):1–11, 1996.

[8] D.J.D. Earn, P. Rohani, and B.T. Grenfell. Persistence, Chaos and Synchrony in Ecology and Epidemiology. *Proc. R. Soc. B*, 265(1390):7–10, 1998.

[9] P. Rohani, D.J.D. Earn, and B.T. Grenfell. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286(5441):968, 1999.

[10] D.J.D. Earn, S.A Levin, and P. Rohani. Coherence and Conservation. *Science*, 290(5495):1360–1364, 2000. doi: 10.1126/science.290.5495.1360.

[11] B. Cazelles, S. Bottani, and L. Stone. Unexpected coherence and conservation. *Proc. R. Soc. B*, 268(1485):2595–2602, 2001.

[12] T.J. Hagenaars, C.A. Donnelly, and N.M. Ferguson. Spatial heterogeneity and the persistence of infectious diseases. *Journal of Theoretical Biology*, 229(3): 349–359, 2004. doi: 10.1016/j.jtbi.2004.04.002.

[13] C.E. Gunning, E. Erhardt, and H.J. Wearing. Incomplete reporting of pre-vaccine era childhood diseases: a case study of observation process variability. *Proc. R. Soc. B*, In Review.

[14] S. Baron, E. Njamkepo, E. Grimprel, P. Begue, J.C. Desenclos, J. Drucker, and N. Guiso. Epidemiology of pertussis in French hospitals in 1993 and 1994: thirty years after a routine use of vaccination. *Pediatr. Infect. Dis. J.*, 17(5):412–418, 1998.

[15] S. Mattoo and J.D. Cherry. Molecular pathogenesis, epidemiology, and clinical manifestations of respiratory infections due to Bordetella pertussis and other Bordetella subspecies. *Clin. Microbiol. Rev.*, 18(2):326–382, 2005. doi: 10.1128/CMR.18.2.326382.2005.

[16] M.S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 4, pages 81–109. University of California Press Berkeley, 1956.

[17] R.M. Anderson and R.M. May. *Infectious diseases of humans*. Oxford University Press Oxford, 1991.

[18] M.J. Ferrari, R.F. Grais, N. Bharti, A.J.K. Conlan, O.N. Bjørnstad, L.J. Wolfson, P.J. Guerin, A. Djibo, and B.T. Grenfell. The dynamics of measles in sub-Saharan Africa. *Nature*, 451(7179):679–684, 2008.

[19] R.M. Anderson and R.M. May. Directly transmitted infectious diseases: control by vaccination. *Science*, 215(4536):1053–1060, 1982.

[20] M.S. Bartlett. Measles periodicity and community size. *J R Stat Soc Ser A*, 120(1):48–70, 1957.

[21] R.M. Anderson, B.T. Grenfell, and R.M. May. Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. *J Hyg (Lond)*, 93(03):587–608, 1984.

[22] M.C. Gomes, J.J. Gomes, and A.C. Paulo. Diphtheria, pertussis, and measles in Portugal before and after mass vaccination: A time series analysis. *Eur. J. Epidemiol.*, 15(9):791–798, 1999.

[23] D.J.D. Earn, P. Rohani, B.M. Bolker, and B.T. Grenfell. A simple model for complex dynamical transitions in epidemics. *Science*, 287(5453):667, 2000.

[24] C.T. Bauch and D.J.D. Earn. Transients and attractors in epidemics. *Proc. R. Soc. B*, 270(1524):1573–1578, 2003.

[25] L. Stone, R. Olinky, and A. Huppert. Seasonal dynamics of recurrent epidemics. *Nature*, 446(7135):533–536, 2007.

[26] H. Broutin, C. Viboud, B.T. Grenfell, M.A. Miller, and P. Rohani. Impact of vaccination and birth rate on the epidemiology of pertussis: a comparative study in 64 countries. *Proc. R. Soc. B*, pages 1–7, 2010. doi: 10.1098/rspb.2010.0994.

[27] I. Nåsell. On the time to extinction in recurrent epidemics. *Proc. R. Soc. B*, 61 (2):309–330, 1999.

[28] H. Andersson and T. Britton. Stochastic epidemics in dynamic populations: quasi-stationarity and extinction. *J Math Biol*, 41(6):559–580, 2000.

[29] A.L. Lloyd. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theor Popul Biol*, 60(1):59–71, 2001.

[30] H.J. Wearing and P. Rohani. Estimating the duration of pertussis immunity using epidemiological signatures. *PLoS Pathog*, 5(10):e1000647, 2009.

[31] A.J.K. Conlan, P. Rohani, A.L. Lloyd, M. Keeling, and B.T. Grenfell. Resolving the impact of waiting time distributions on the persistence of measles. *J R Soc Interface*, 7(45):623, 2010.

[32] M.J. Ferrari, B.T. Grenfell, and P.M. Strebel. Think globally, act locally: the role of local demographics and vaccination coverage in the dynamic response of measles infection to control. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368 (1623):20120141, 2013.

[33] C.E. Gunning and H.J. Wearing. Probabilistic measures of persistence and extinction in measles (meta)populations. *Ecol. Lett.*, 16:985–994, 2013.

[34] B.T. Grenfell, B.M. Bolker, and A. Kleczkowski. Seasonality and extinction in chaotic metapopulations. *Proc. R. Soc. B*, 259(1354):97–103, 1995.

[35] B.M. Bolker and B.T. Grenfell. Impact of vaccination on the spatial correlation and persistence of measles dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 93(22): 12648–12653, 1996.

[36] M.J. Keeling and P. Rohani. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecol. Lett.*, 5(1):20–29, 2002.

[37] A.W. Park, S. Gubbins, and C.A. Gilligan. Invasion and persistence of plant parasites in a spatially structured host population. *Oikos*, 94(1):162–174, 2001.

[38] B. Grenfell and J. Harwood. (Meta) population dynamics of infectious diseases. *Trends Ecol. Evol. (Amst.)*, 12(10):395–399, 1997.

[39] N.S. Crowcroft, C. Stein, P. Duclos, and M. Birmingham. How best to estimate the global burden of pertussis? *Lancet Infect Dis*, 3(7):413–418, 2003. doi: 10.1016/S1473-3099(03)00669-8.

[40] R.E. Black, S. Cousens, H.L. Johnson, J.E. Lawn, I. Rudan, D.G. Bassani, P. Jha, H. Campbell, C.F. Walker, R. Cibulskis, T. Eisele, L. Liu, and C. Mathers. Global, regional, and national causes of child mortality in 2008: a systematic analysis. *Lancet*, 375(9730):1969–1987, 2010. doi: 10.1016/S0140-6736(10)60549-1.

[41] M.N. Mulders, A.T. Truong, and C.P. Muller. Monitoring of measles elimination using molecular epidemiology. *Vaccine*, 19(17):2245–2249, 2001.

[42] S.L. Katz, J.I. Santos, M.A. Nakamura, M.V. Godoy, P. Kuri, C.A. Lucas, and R.T. Conyer. Measles in Mexico, 1941–2001: Interruption of Endemic Transmission and Lessons Learned. *J. Infect. Dis.*, 189(Supplement 1):S243–S250, 2004. doi: 10.1086/378520.

[43] E. Kaliner, J. Moran-Gilad, I. Grotto, E. Somekh, E. Kopel, M. Gdalevich, E. Shimron, Y. Amikam, A. Leventhal, B. Lev, and R. Gamzu. Silent reintroduction of wild-type poliovirus to Israel, 2013–risk communication challenges in an argumentative atmosphere. *Euro Surveill*, 19(7):207030, 2014.

[44] J.O. Lloyd-Smith, P.C. Cross, C.J. Briggs, M. Daugherty, W.M. Getz, J. Latto, M.S. Sanchez, A.B. Smith, and A. Swei. Should we expect population thresholds for wildlife disease? *Trends Ecol. Evol.*, 20(9):511–519, 2005.

[45] W.H. McNeill. *Plagues and Peoples*. Anchor, 1977.

[46] A.J.K. Conlan and B.T. Grenfell. Seasonality and the persistence and invasion of measles. *Proc. R. Soc. B*, 274(1614):1133–1141, 2007.

[47] A.R. McLean and R.M. Anderson. Measles in developing countries. Part II. The predicted impact of mass vaccination. *Epidem. Inf.*, 100:419–442, 1988.

[48] J.S. Lavine, A.A. King, and O.N. Bjørnstad. Natural immune boosting in pertussis dynamics and the potential for long-term vaccine failure. *Proc. Natl. Acad. Sci. U.S.A.*, 108(17):7259–7264, 2011.

[49] L.L. Rodriguez and C.G. Gay. Development of Vaccines Toward the Global Control and Eradication of Foot-and-mouth Disease. *Expert Rev Vaccines*, 10 (3):377–387, 2011.

[50] A Uttenthal, S. Parida, T.B. Rasmussen, D.J. Paton, B. Haas, and W.G. Dundon. Strategies for differentiating infection in vaccinated animals (DIVA) for foot-and-mouth disease, classical swine fever and avian influenza. *Expert Rev Vaccines*, 9(1):73–87, 2010. doi: 10.1586/erv.09.130.

[51] M.L. Gennaro. Immunologic diagnosis of tuberculosis. *Clin. Infect. Dis.*, 30 (Supplement 3):S243–S246, 2000. doi: 10.1086/313868. URL `http://cid.oxfordjournals.org/content/30/Supplement_3/S243.abstract`.

[52] A. Story, S. Murad, M. Verheyen, W. Roberts, and A.C. Hayward. Tuberculosis in London-the importance of homelessness, problem drug use and prison. *Thorax*, 62(8):667–671, 2007.

[53] S.E. Dorman. New diagnostic tests for tuberculosis: Bench, bedside, and beyond. *Clin. Infect. Dis.*, 50(Supplement 3):S173–S177, 2010. doi: 10.1086/651488. URL `http://cid.oxfordjournals.org/content/50/Supplement_3/S173.abstract`.

[54] N.R. Gandhi, P. Nunn, K. Dheda, H.S. Schaaf, M. Zignol, D. Van Soolingen, P. Jensen, and J. Bayona. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet*, 375(9728):1830–1843, 2010.

[55] S. Sarkar, X.L. Tang, D. Das, J.S. Spencer, T.L. Lowary, and M.R. Suresh. A Bispecific Antibody Based Assay Shows Potential for Detecting Tuberculosis in Resource Constrained Laboratory Settings. *PLoS ONE*, 7(2):e32340, 02 2012. doi: 10.1371/journal.pone.0032340. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0032340`.

[56] H. Noedl, D. Socheat, and W. Satimai. Artemisinin-Resistant Malaria in Asia. *N. Engl. J. Med.*, 361(5):540–541, 2009. doi: 10.1056/NEJMc0900231. URL `http://www.nejm.org/doi/full/10.1056/NEJMc0900231`. PMID: 19641219.

[57] T. Brown, L.S. Smith, E.K.S. Oo, K. Shawng, T.J. Lee, D. Sullivan, C. Beyrer, and A.K. Richards. Molecular surveillance for drug-resistant Plasmodium falciparum in clinical and subclinical populations from three border regions of Burma/Myanmar: cross-sectional data and a systematic review of resistance studies. *Malar. J.*, 11:333, 2012. doi: 10.1186/1475-2875-11-333.

[58] S. Hoyer, S. Nguon, S. Kim, N. Habib, N. Khim, S. Sum, E. Christophel, S. Bjorge, A. Thomson, S. Kheng, et al. Focused Screening and Treatment (FSAT): a PCR-based Strategy to Detect Malaria Parasite Carriers and Contain Drug Resistant P. falciparum, Pailin, Cambodia. *PloS ONE*, 7(10):e45797, 2012. doi: 10.1371/journal.pone.0045797.

[59] B. Wang, S. Han, C. Cho, J. Han, Y. Cheng, S. Lee, G.N.L. Galappaththy, K. Thimasarn, M.T. Soe, H.W. Oo, et al. Comparison of Microscopy, Nested-PCR, and Real-Time-PCR Assays Using High-Throughput Screening of Pooled Samples for Diagnosis of Malaria in Asymptomatic Carriers from Areas of Endemicity in Myanmar. *J. Clin. Microbiol.*, 52(6):1838–1845, 2014. doi: 10.1128/JCM.03615-13.

[60] A. Levitt, O.M. Diop, R.H. Tangermann, F. Paladin, J.B. Kamgang, C.C. Burns, P.J. Chenoweth, A. Goel, and S.G. Wassilak. Surveillance systems to

track progress toward global polio eradication-worldwide, 2012-2013. *MMWR Morb. Mortal. Wkly. Rep.*, 63(16):356–361, 2014.

[61] C. Willyard. Polio: The eradication endgame. *Nature*, 507(7490):S14–S15, 2014.

[62] R.B. Aylward and A. Alwan. Polio in Syria. *Lancet*, 383:489–491, 2014.

4.9 Supplemental Information

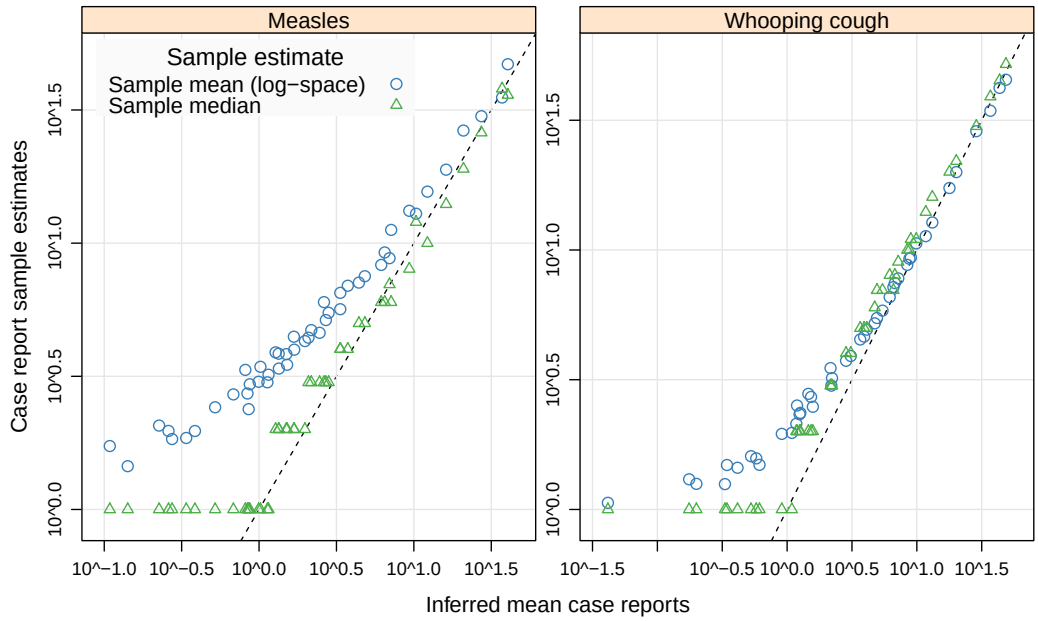

Figure 4.4. Log-log scaling of inferred mean case reports with sample estimates, including the sample median of case reports and the unlogged sample mean of log case reports. For both sample estimates, zeros are treated as 1 case report. Exclusion of zeros artificially increases the sample mean, and does not affect non-zero medians. The dashed black line shows the 1-to-1 line.

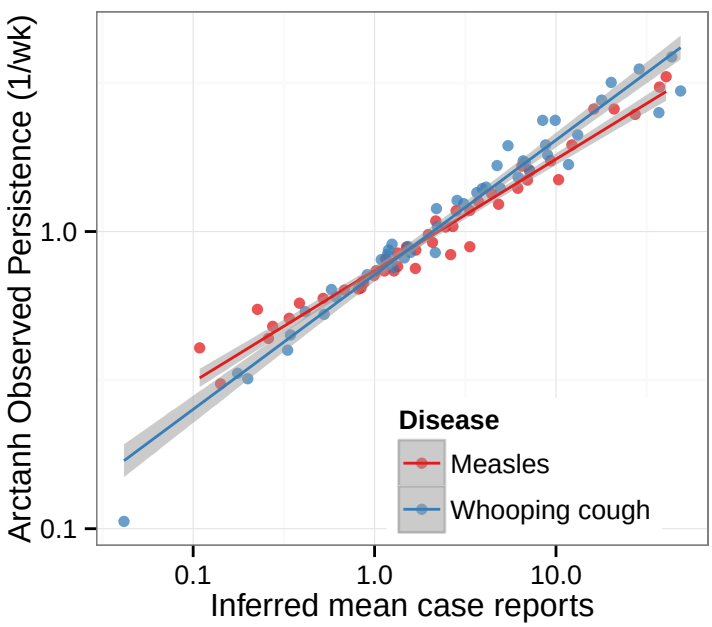

Figure 4.5. Observed persistence (Fisher transformed = arctanh) versus inferred mean case reports. Linear models are also shown, which were fit in log-log space. Adjusted R^2 for linear models: Measles = 0.96; Whooping cough = 0.96.

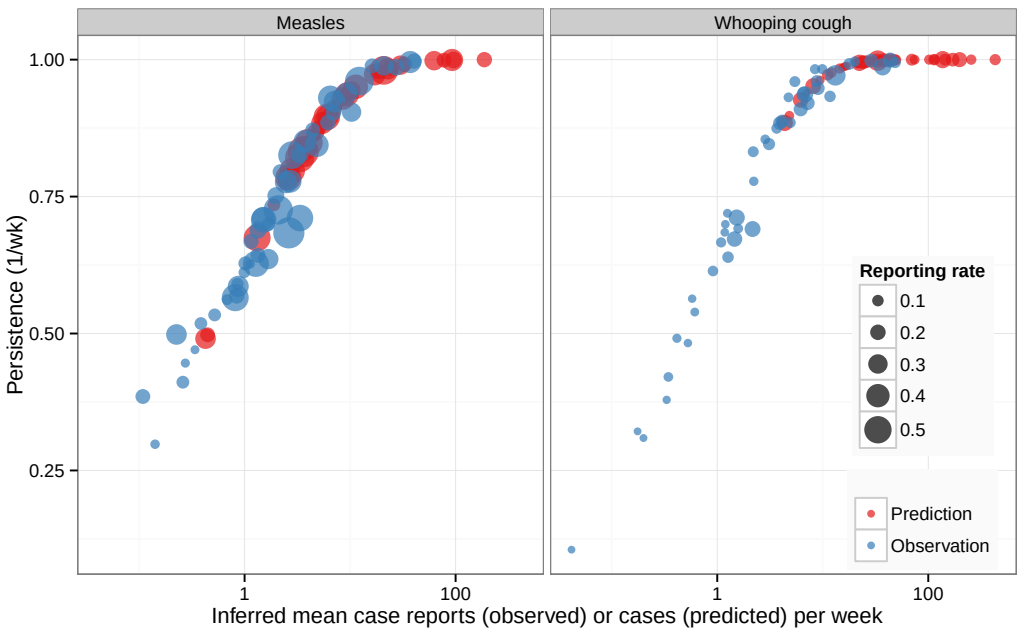

Figure 4.6. Observed (predicted) persistence versus observed (predicted) incidence. The meaning of the x-axis changes between groups: observed persistence is plotted against inferred mean case reports, while predicted persistence is plotted against inferred mean cases at full reporting. Correction for incomplete reporting transforms case reports (observed incidence) to cases (predicted incidence). The linear model for each disease then predicts persistence. The overall motion of a location's point is rightwards (correction for incomplete reporting) and upwards (new model prediction). Thus uncertainty in predicted incidence includes uncertainty in both reporting probability and linear model predictions.

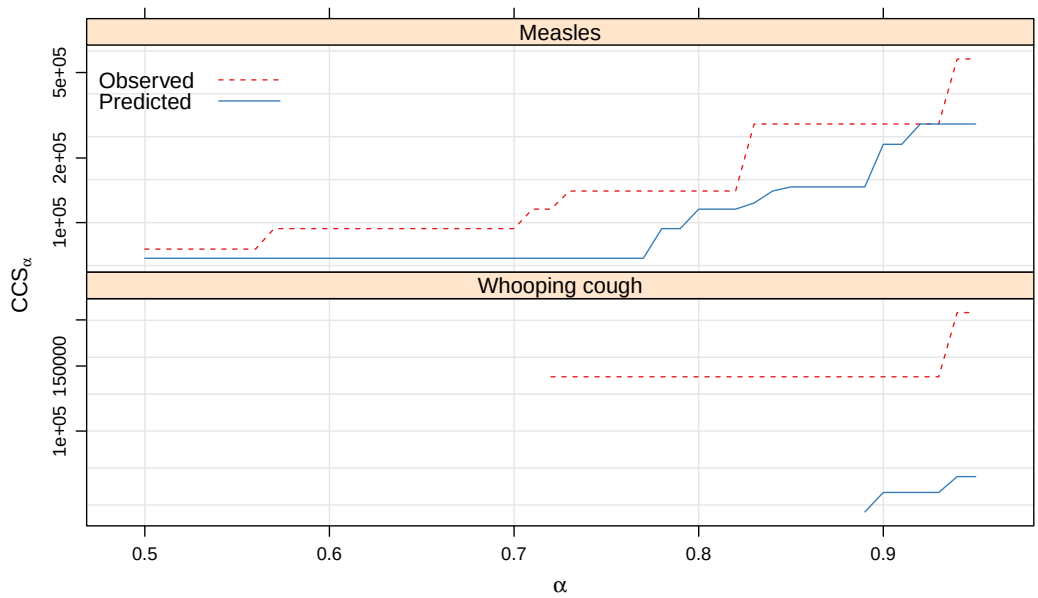

Figure 4.7. Empirical estimates of $CCS_\alpha := min(\text{Population})|P > \alpha$ for a range of α, using P_o (Observed) and $\tilde{P}$ (Predicted). Whooping cough is predicted to persist in all cities during more than 85% of sampled weeks. The minimum sampled population size is 38 thousand; CCS estimates equal to this minimum city size lack any meaning, and are excluded.