

# LEARNING FROM NEIGHBORS WITH PHIBP: PREDICTING INFECTIOUS DISEASE DYNAMICS IN DATA-SPARSE ENVIRONMENTS

BY EDWIN FONG<sup>1,a</sup>, LANCELOT F. JAMES<sup>2,b</sup>, AND JUHO LEE<sup>3,c</sup>

<sup>1</sup>DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, HKU, <sup>a</sup>[CHEFONG@HKU.HK](mailto:CHEFONG@HKU.HK)

<sup>2</sup>DEPARTMENT OF ISOM, HKUST, <sup>b</sup>[LANCELOT@UST.HK](mailto:LANCELOT@UST.HK)

<sup>3</sup>THE GRADUATE SCHOOL OF AI, KAIST, <sup>c</sup>[JUHOLEE@KAIST.AC.KR](mailto:JUHOLEE@KAIST.AC.KR)

Modeling sparse count data, which arise across numerous scientific fields, presents significant statistical challenges. This chapter addresses these challenges in the context of infectious disease prediction, with a focus on predicting outbreaks in geographic regions that have historically reported zero cases. To this end, we present the detailed computational framework and experimental application of the Poisson Hierarchical Indian Buffet Process (PHIBP) [5], a Bayesian machine learning model with demonstrated success in handling sparse count data in microbiome and ecological studies. The PHIBP's architecture, grounded in the concept of absolute abundance, systematically borrows statistical strength from related regions and circumvents the known sensitivities of relative-rate methods to zero counts. Through a series of experiments on infectious disease data, we show that this principled approach provides a robust foundation for generating coherent predictive distributions and for the effective use of comparative measures such as alpha and beta diversity. The chapter's emphasis on algorithmic implementation and experimental results confirms that this unified framework delivers both accurate outbreak predictions and meaningful epidemiological insights in data-sparse environments.

**1. Introduction.** Predicting disease prevalence in regions that have not yet reported cases—despite clear outbreaks in neighboring areas—is a problem most people now recognize from lived experience: when will it reach us, and how severe might it be? Turning this intuitive concern into reliable forecasts is a statistically challenging task. Historically, one can relate this to classes of count-based models for both realized abundance (counts) and prediction of unseen numbers and types of species in the classic work of [2]. There Fisher, in answer to Corbett’s question about the expected number of distinct species of butterflies to be seen in a future sample, developed the logarithmic series distribution to describe the prevalence of individually observed species. This then translates into the remarkable work of [6], which empirically verifies the Negative Binomial distribution as a viable model for soil microorganism counts in the form of a mixed Poisson variable: the framework in [6] provides an empirical justification for a model in which the number of colonies is distributed according to a Poisson distribution, while the bacterial counts within each colony independently follow the logarithmic series distribution introduced by [2].

Naturally, while many core elements of those works translate to modern datasets, they do not capture the notion of sharing information across groups, nor matters

---

*MSC2020 subject classifications:* Primary 60G09, 62F15; secondary 60G57, 62P10, 60C05.

*Keywords and phrases:* Bayesian nonparametrics, Bayesian statistical machine learning, Hierarchical Indian Buffet Process, Microbiome species sampling models, Microbiome unseen species problems.

of count sparsity. In particular, while the prevalence of zeros is common in modern datasets, we further highlight the notion of sparse co-occurrence, where samples across groups do not share many common species. These challenges, along with modeling complex multivariate count distributions, are all addressed by the Poisson Hierarchical Indian Buffet Process (PHIBP) [5] as applied to complex microbiome sampling models. Here, we focus on its ability to pool information across groups to handle zero observations, leading to credible predictive and inferential modeling for infectious disease prediction. As the theory of PHIBP has been detailed previously, our focus is on methodology and computational inference within the present context of disease prediction.

**2. The PHIBP model.** We consider  $J$  related regions (e.g. counties),  $j \in [J]$ , each with  $M_j$  replicated count samples (e.g. years),  $i \in [M_j]$ . A sample has the form

$$Z_j^{(i)} = \sum_{l \geq 1} N_{j,l}^{(i)} \delta_{Y_l},$$

where  $Y_l$  indexes a global catalogue of disease types and  $N_{j,l}^{(i)} \in \{0, 1, 2, \dots\}$  is the count of disease  $Y_l$  in sample  $i$  from region  $j$ . The PHIBP specifies a hierarchical prior on the latent mean abundance rates that generate these counts. At the top level, we draw a global completely random measure

$$B_0 = \sum_{l \geq 1} \lambda_l \delta_{Y_l} \sim \text{CRM}(\tau_0, F_0),$$

where  $\tau_0$  is a Lévy measure on  $(0, \infty)$  and  $F_0$  a non-atomic probability measure on a Polish space  $\Omega$  of disease labels. The jumps  $(\lambda_l)$  are interpreted as global mean rates of the diseases ( $Y_l$ ) across all regions.

Given  $B_0$ , each region  $j$  has a local CRM

$$B_j | B_0 \sim \text{CRM}(\tau_j, B_0), \quad j \in [J],$$

with Lévy density  $\tau_j$  on  $(0, \infty)$ . Conditionally on  $B_0 = \sum_l \lambda_l \delta_{Y_l}$  we may write

$$B_j \stackrel{d}{=} \sum_{l \geq 1} \sigma_{j,l}(\lambda_l) \delta_{Y_l}, \quad \sigma_{j,l}(\lambda_l) \stackrel{d}{=} \sigma_j(\lambda_l),$$

where  $(\sigma_j(t) : t \geq 0)$  is a subordinator with Lévy density  $\tau_j$ , satisfying for  $s < t$ ,  $\sigma_j(t) - \sigma_j(s) \stackrel{d}{=} \sigma_j(t-s)$  independent of  $\sigma_j(s)$ . Thus  $\lambda_l$  is the global rate of disease  $Y_l$ , and  $\sigma_{j,l}(\lambda_l)$  is its local mean rate in region  $j$ .

Throughout, as in [5], we denote by  $\psi_j(\cdot)$  and  $\Psi_0(\cdot)$  the Laplace exponents associated with the Lévy measures  $\tau_j$  and  $\tau_0$ , respectively. In particular, for each  $j$  and exposure  $\gamma_j > 0$ ,  $\mathbb{E}[e^{-\gamma_j \sigma_j(\lambda)}] = e^{-\lambda \psi_j(\gamma_j)}$ , where

$$\psi_j(\gamma_j) = \int_{(0,\infty)} (1 - e^{-\gamma_j s}) \tau_j(ds),$$

and similarly

$$\Psi_0(\gamma_0) = \int_{(0,\infty)} (1 - e^{-\gamma_0 \lambda}) \tau_0(d\lambda), \quad \gamma_0 \geq 0.$$

where we assume that  $\int_0^\infty \min(s, 1) \tau_j(s) ds < \infty$  ensuring the finiteness of  $B_j(\Omega)$  for  $j = 0, \dots, J$ . We will only use these via evaluations such as  $\psi_j(\sum_{i=1}^{M_j} \gamma_{i,j})$  and  $\Psi_0(\sum_{j=1}^J \psi_j(\cdot))$ , exactly as in the PHIBP construction of [5]. Furthermore setting

$\psi_j^{(c)}(\gamma_j) = \int_0^\infty s^c e^{-\gamma_j s} \tau_j(s) ds$  we have the mixed truncated Poisson distributions with law denoted as MtP( $\tau_j, \gamma_j$ ) and probability mass function for  $c = 1, 2, \dots$ , given by

$$\frac{\gamma_j^c \psi_j^{(c)}(\gamma_j)}{\psi_j(\gamma_j) c!}$$

The PHIBP ( $Z_j^{(i)}, i \in [M_j], j \in [J]$ ) is specified as a mixed Poisson processes driven by the local CRMs. For each  $j$  and  $i \in [M_j]$ ,

$$(2.1) \quad Z_j^{(i)} | B_j \stackrel{\text{ind}}{\sim} \text{PoiP}(\gamma_{i,j} B_j),$$

where  $\gamma_{i,j} > 0$  is a sample-specific exposure. Using the atomic form of  $B_j$ , this means

$$(2.2) \quad Z_j^{(i)} \stackrel{d}{=} \sum_{l \geq 1} \mathcal{P}_{j,l}^{(i)}(\gamma_{i,j} \sigma_{j,l}(\lambda_l)) \delta_{Y_l},$$

with conditionally independent Poisson counts

$$\mathcal{P}_{j,l}^{(i)}(\gamma_{i,j} \sigma_{j,l}(\lambda_l)) \sim \text{Poisson}(\gamma_{i,j} \sigma_{j,l}(\lambda_l)).$$

This is the basic mixed Poisson representation: each observed count is Poisson with a random intensity given by the exposure  $\gamma_{i,j}$  times the local rate  $\sigma_{j,l}(\lambda_l)$ .

Summing over samples within region  $j$  gives

$$\sum_{i=1}^{M_j} Z_j^{(i)} \Big| B_j \sim \text{PoiP}\left(\left(\sum_{i=1}^{M_j} \gamma_{i,j}\right) B_j\right)$$

and induces the usual quantity

$$\psi_j\left(\sum_{i=1}^{M_j} \gamma_{i,j}\right) = \int_{(0,\infty)} \left(1 - e^{-s \sum_{i=1}^{M_j} \gamma_{i,j}}\right) \tau_j(ds),$$

which governs the rate of species (atoms) observed in region  $j$  after aggregating over its  $M_j$  samples. Writing  $N_{j,l} := \sum_{i=1}^{M_j} N_{j,l}^{(i)}$  for the total count of disease  $Y_l$  in region  $j$ , we have

$$N_{j,l} | (\sigma_{j,l}(\lambda_l))_{j,l} \sim \text{Poisson}\left(\left(\sum_{i=1}^{M_j} \gamma_{i,j}\right) \sigma_{j,l}(\lambda_l)\right),$$

independently over  $(j, l)$ . The key result (Theorem 3.1 of [5]) shows that the joint law of the summed processes

$$(2.3) \quad \left(\sum_{i=1}^{M_j} Z_j^{(i)}, j \in [J]\right) \stackrel{d}{=} \left(\sum_{\ell=1}^{\varphi} \tilde{N}_{j,\ell} \delta_{\tilde{Y}_\ell}, j \in [J]\right) \stackrel{d}{=} \left(\sum_{\ell=1}^{\varphi} \left[\sum_{k=1}^{X_{j,\ell}} C_{j,k,\ell}\right] \delta_{\tilde{Y}_\ell}, j \in [J]\right).$$

admits an exact compound Poisson representation in terms of a random number of latent (sub)-species level clusters and their counts. Here  $\varphi$  is the random number of distinct species that appear across all regions, and has a Poisson distribution with mean  $\Psi_0(\sum_{j=1}^J \psi_j(\sum_{i=1}^{M_j} \gamma_{i,j}))$ . For each observed species  $\ell$ , there is a global posterior mean rate  $H_\ell > 0$  and a corresponding total number of OTU-level clusters  $X_\ell := \sum_{j=1}^J X_{j,\ell}$ ; jointly,  $(H_\ell, X_\ell)$  has the mixed Poisson–MtP structure of [5], with  $X_\ell \sim \text{MtP}(\tau_0, \sum_{j=1}^J \psi_j(\sum_{i=1}^{M_j} \gamma_{i,j}))$  and  $H_\ell$  conditionally distributed with density proportional to  $\lambda^{X_\ell} \exp\{-\lambda \sum_{j=1}^J \psi_j(\sum_{i=1}^{M_j} \gamma_{i,j})\} \tau_0(\lambda)$ . Given  $X_\ell = x_\ell$ , the allocation of OTU clusters across regions is multinomial,  $(X_{1,\ell}, \dots, X_{J,\ell}) | X_\ell = x_\ell \sim \text{Multinomial}(x_\ell; q_1, \dots, q_J)$ ,

with  $q_j \propto \psi_j(\sum_{i=1}^{M_j} \gamma_{i,j})$ . Finally, given these allocations and the local Lévy measures  $\tau_j$ , the OTU-level cluster sizes  $C_{j,k,\ell}$  are independent mixed truncated Poisson variables  $C_{j,k,\ell} \sim \text{MtP}(\tau_j, \sum_{i=1}^{M_j} \gamma_{i,j})$ , so that each regional total  $\tilde{N}_{j,\ell} = \sum_{k=1}^{X_{j,\ell}} C_{j,k,\ell}$  is a sum of i.i.d. MtP components driven by the global rate  $H_\ell$ .

**2.1. Predicting the unseen and alpha/beta diversities.** We now recount some of the novel developments in [5] that will figure prominently in our analysis. We will use this compound Poisson form, together with the posterior local rates  $\tilde{\sigma}_{j,l}(H_l)$ , to define and compute our Bayesian alpha- and beta-diversity measures and disease prediction functionals. Specifically the representation in (2.3) indicates there is a decomposition of the pairs of rates and species labels  $(\lambda_l, Y_l)_{l \geq 1}$  in terms of unobserved species and corresponding rates  $(\lambda'_l, Y'_l)_{l \geq 1}$  and the  $\varphi$  observed species  $\tilde{Y}_\ell$  and corresponding rates  $H_\ell$  for  $\ell \in [\varphi]$ . The localized posterior rates  $\tilde{\sigma}_{j,l}(H_l)$  allow us to construct novel Bayesian analogues of classical  $\alpha$ - and  $\beta$ -diversity metrics that account for uncertainty and unobserved species.

Within this framework, among other possibilities, we define alpha-diversity (within-group diversity) via the Shannon entropy, which is the random variable:

$$(2.4) \quad \mathcal{D}_j := - \sum_{l=1}^{\varphi} \frac{\tilde{\sigma}_{j,l}(H_l)}{\sum_{t=1}^{\varphi} \tilde{\sigma}_{j,t}(H_t)} \log \left( \frac{\tilde{\sigma}_{j,l}(H_l)}{\sum_{t=1}^{\varphi} \tilde{\sigma}_{j,t}(H_t)} \right).$$

For beta-diversity (between-group diversity), we define a dissimilarity based on the Bray-Curtis index, which is the random variable:

$$(2.5) \quad \mathcal{B}_{j,v} := \frac{\sum_{l=1}^{\varphi} |\tilde{\sigma}_{j,l}(H_l) - \tilde{\sigma}_{v,l}(H_l)|}{\sum_{l=1}^{\varphi} (\tilde{\sigma}_{j,l}(H_l) + \tilde{\sigma}_{v,l}(H_l))}, \quad j \neq v \in [J].$$

Our constructs while new are in the form of more classical measures of diversity as detailed in [9, 10]. In ecological and epidemiological settings, *alpha-diversity* measures the within-region richness and evenness of types—in our case, how many diseases circulate in a county and how evenly their burdens are distributed—while *beta-diversity* quantifies between-region dissimilarity, capturing how disease profiles differ across counties or how transmission patterns diverge geographically. Classically, these metrics are computed from observed proportions, but such compositional methods break down in the presence of many zeros, causing artificial inflation of dissimilarity when low-prevalence types are simply unobserved. The PHIBP framework replaces fixed proportions with *latent abundance rates*, producing alpha- and beta-diversities that properly distinguish sparse detection from true absence. Moreover, these quantities are Bayesian random variables, yielding full posterior distributions that reflect uncertainty—critical when data are sparse or uneven across regions—rather than single point estimates. This same decomposition underlies PHIBP’s approach to the unseen-disease problem, enabling principled prediction of diseases not yet observed but likely to appear under continued sampling. In our previous PHIBP experiments on the Dorado Outcrop microbiome data, which is based on data originally investigated by [7, 10], this Bayesian, rate-based treatment proved essential: the model captured rare shared taxa, avoided pseudocount artifacts, and produced more stable alpha- and beta-diversity estimates than compositional alternatives. The same advantages carry over directly to the disease-prediction setting studied here.

**3. Infectious disease dataset.** The dataset of focus in this chapter is the ‘Infectious Diseases by Disease, County, Year, Sex’ dataset collected by the California Department of Public Health [8]. The dataset contains the prevalence of a selection of communicable infectious diseases across all 58 counties within California from 2001 - 2023 obtained through local health providers and laboratories. Due to factors such as incomplete reporting by healthcare providers or lack of access to healthcare, the reported prevalence of many diseases is likely an underestimate of the true disease prevalence, so we do not expect underlying rates to actually be zero. A key property of the dataset is thus count sparsity, where there are multiple regions with no reported counts of specific diseases. For example, in 2023, there were 144 reported cases of Listeriosis, but these cases were all confined to occur in only 28 out of 58 counties. As the diseases are communicable, there is also the expectation of some prevalence correlations occurring due to geographical vicinity, which we will quantify through the aforementioned beta-diversity and visualize in a geographical heatmap.

The main goal of our analysis is to predict disease counts in each county, paying careful attention to county-disease pairs that have 0 reported counts. We will then proceed to validate these predictions in a held-out test set to highlight the advantages of information borrowing between counties. A more exploratory goal is to evaluate whether or not geographical vicinity is a driver in correlation of the infectious disease landscape. To carry out the latter task, we will compute the estimated beta-diversities for different reference counties and visualize the values on a geographical heatmap.

**3.1. Experimental details.** For our analysis, we begin by filtering out highly common infectious diseases to focus our study on rarer diseases. This includes disease such as Salmonellosis and Campylobacteriosis which have a maximum count of over 1000 in a single county, leaving us with 36 remaining infectious diseases. We will then fit the PHIBP model to the years 2001 - 2014, treating the remaining years 2015-2023 as the held-out test dataset.

For the PHIBP model, we specify both  $\tau_0$  and  $\tau_j$  as Lévy measures of a Generalized Gamma (GG) subordinator. Each measure is parameterized as

$$\tau_j(\lambda) = \frac{\theta_j}{\Gamma(1 - \alpha_j)} \lambda^{-1 - \alpha_j} e^{-\lambda} \text{ for } j \in \{0\} \cup [J],$$

where  $\theta_j > 0$  and  $\alpha_j \in [0, 1)$  for all  $j \in 0 \cup [J]$ . The parameter  $\alpha_j$  controls the degree of population diversity: larger values of  $\alpha_j$  induce richer latent structures, resulting in a more diverse collection of disease types appearing within or across counties. A notable special case arises when  $\alpha_j = 0$ , in which case the GG subordinator reduces to the Gamma (GA) subordinator. Since GA subordinators exhibit distinct asymptotic behavior, we treat GA as a separate baseline model and compare it to the GG specification with  $\alpha_j > 0$ .

For posterior inference, we run three independent MCMC chains for both GG and GA versions of the PHIBP model. Each chain is executed for 40,000 iterations, with the first 20,000 iterations discarded as burn-in. We retain samples using a thinning interval of 10 to reduce autocorrelation. Additional implementation details—including prior specification, sampling scheme, and prediction procedures —are provided in [5].

**3.2. Results.** Figure 1 (left) compares the GG and GA models in terms of test log-likelihood, computed following [5], and shows that GG clearly outperforms GA, likely due to its improved ability to capture rare diseases. Figure 1 (middle) and (right) present the posterior distributions of  $\alpha_0$  (global) and  $\alpha_{37}$  (San Francisco), along with

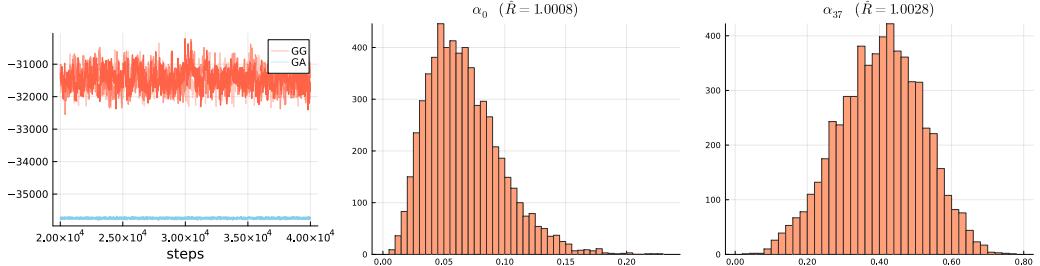


Fig 1: (Left) test log-likelihoods of GG and GA PHIBP models. (Middle) posterior distribution of  $\alpha_0$ . (Right) posterior distribution of  $\alpha_{37}$  (corresponding to San Francisco).

their  $\hat{R}$  values [1, 3], where values close to 1 indicate good mixing. The posterior for  $\alpha_0$  is concentrated near zero, suggesting limited diversity of rare diseases overall, whereas  $\alpha_{37}$  is centered around approximately 0.4, indicating substantially greater rare-disease diversity within San Francisco.

An appealing feature of the PHIBP model is its ability to predict novel diseases, i.e., to assign positive counts to diseases that were never observed for a given county in the training data. To illustrate this capability, and to contrast the behavior of GG and GA variants, we identify all (county, disease) pairs with zero training-set counts and examine their predicted counts in the test set. Because there are several hundred such pairs, we subsample 12 and plot the predicted counts from the GG and GA models against the true test-set counts in Figure 2. As shown, both models assign nonzero predictions to unseen diseases—an indication of information sharing across groups—but GG typically produces higher estimates that more closely match the observed counts, consistent with its diversity-promoting properties. Similar behavior was reported in the microbiome experiments of [5].

We now proceed to investigate posterior inference on the alpha/beta diversities as a function of geographical location. As beta-diversity is a measure of between-group diversity, we will investigate the beta-diversity for GG with respect to a reference county, which can be interpreted as a single row in a county by county beta-diversity matrix. Figure 3 presents two geographical heatmaps of the counties in California, where the shade represents the posterior mean of the beta-diversities with respect to two reference counties, namely Del Norte (left) and San Diego (right). The darker shade indicates a low beta-diversity, that is similarity in infectious disease profile, and the reference counties are in dark blue as they have 0 beta-diversity when compared to themselves. We visually see a clear trend: for counties near the reference counties, the beta-diversity tends to be lowest in the surrounding regions, with maximum beta-diversity as we move far away from the reference county. This is an intuitive finding, given that infectious diseases are transmissible between people, and geographic vicinity is likely a large driver in this transmission.

Figure 4 presents a similar geographical heatmap for alpha-diversities in each county, where we now compare GG and GA. We again see a geographical clustering in the alpha-diversity in Figure 4, where southern counties tend to exhibit higher alpha-diversity indicating a richer disease profile. We also see in Figure 4 that GG tends to display slightly higher levels of alpha-diversity as expected, which is in agreement with the discussion in Section 3.1 and the results of [5]. Finally, Figure 5 (left) conveys the posterior uncertainty in alpha-diversities, namely the precision of the posterior distribution over the alpha-diversities. We see a direct connection of posterior uncertainty to the average

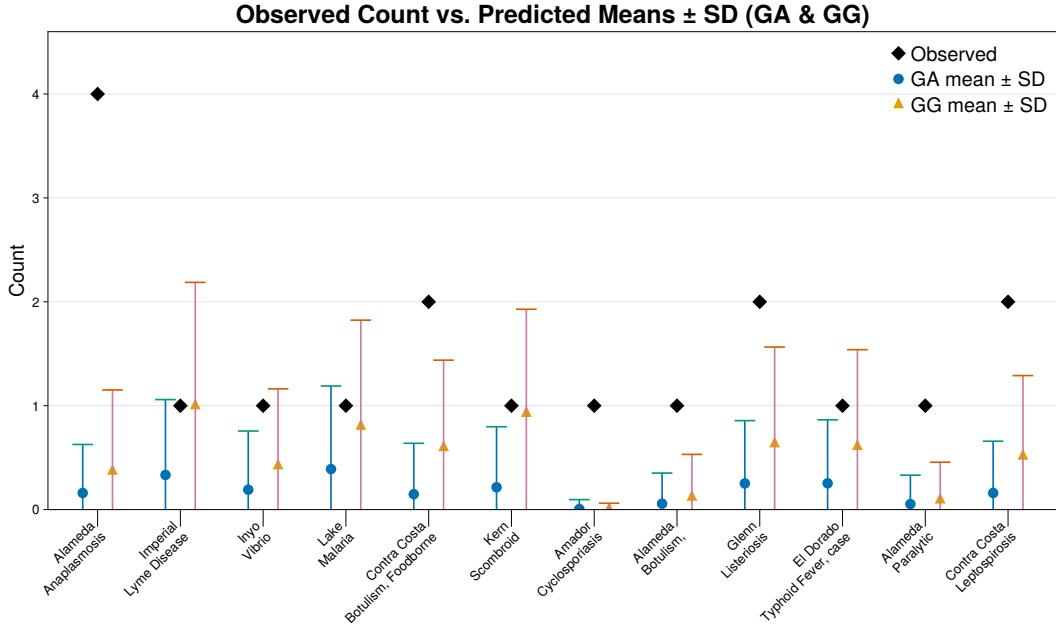


Fig 2: Predicted vs test data counts for county-disease pairs with 0 counts in training dataset.

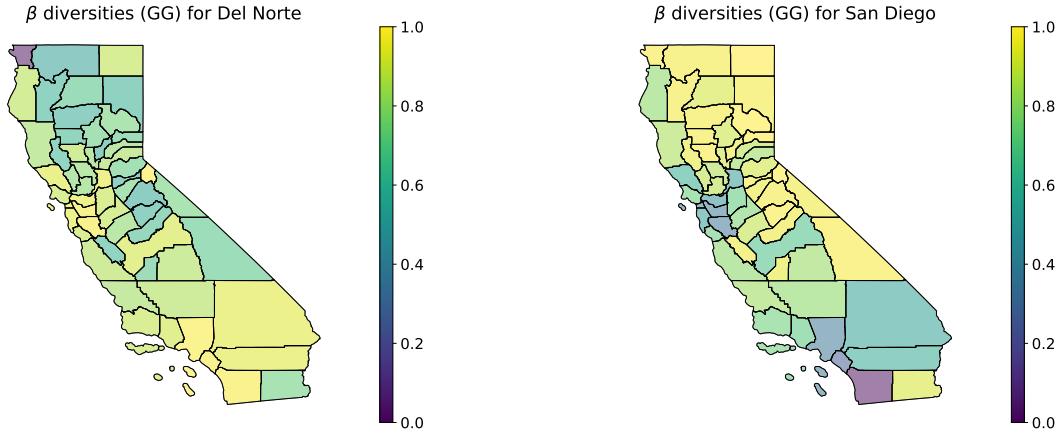


Fig 3: Heatmap of posterior mean of pairwise beta-diversities for GG with reference counties (in dark blue) as Del Norte (left) and San Diego (right).

county population between 2001-2014 plotted in Figure 5 (right), which accurately depicts largest posterior precision for counties with the greatest population such as Los Angeles.

**4. Extensions: From Local Borrowing to Global Architectures.** The PHIBP framework demonstrated here for infectious disease prediction represents a specialized instance of a more fundamental mathematical architecture developed in [4], where  $Z$  represents a coarse process and  $I$  a coupled fine process, both embedded within a four-component simultaneous structural duality system. While our focus has been on

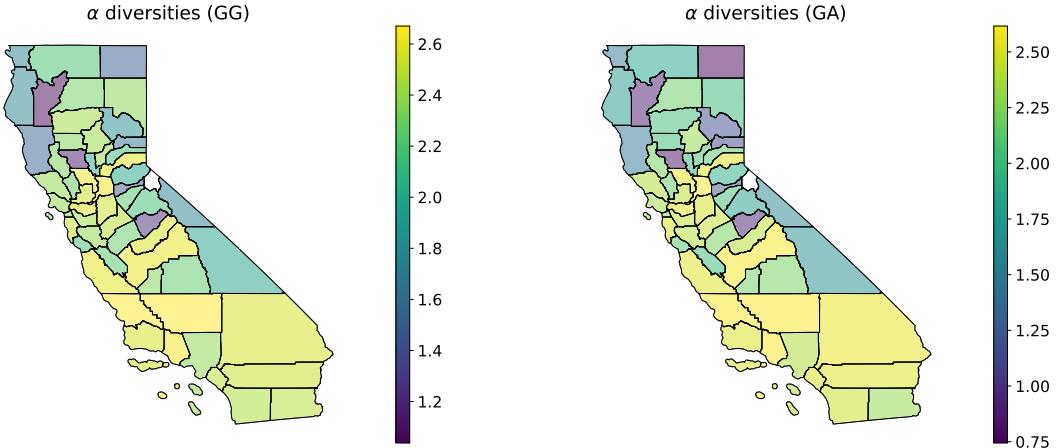


Fig 4: Heatmap of posterior mean of alpha-diversities for GG (left) and GA (right).

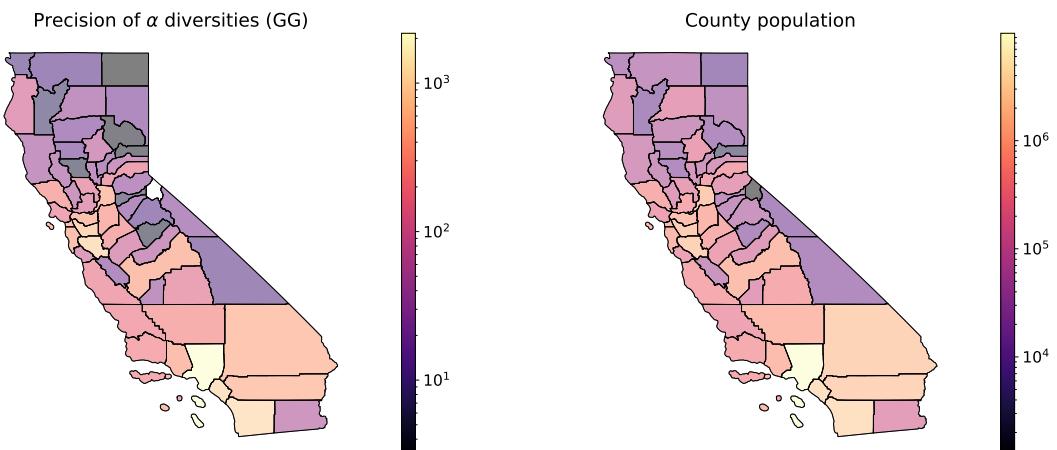


Fig 5: Heatmap of posterior precision of alpha-diversities for GG (left) and average county population from 2001-2014 (right); note the logarithmic scale.

the practical challenge of predicting disease outbreaks from sparse data, the underlying construction connects to deeper principles of information sharing and structural duality—a concept introduced in that work—in hierarchical systems.

**4.1. The Architecture of Information Borrowing.** The success of PHIBP in predicting unseen diseases and capturing geographic clustering in diversity measures reflects a specific realization of what is termed “cloud duality” in [4]—the principle that hierarchical borrowing of information corresponds to reversible coagulation-fragmentation operations on the underlying point processes.

In our disease application, this manifests as the model’s ability to:

- Fragment the global disease pool into county-specific patterns (via the compound Poisson decomposition in equation (2.3))
- Coagulate local observations to inform global patterns (through the hierarchical random measure structure)

This bidirectional flow—visible in how San Francisco’s higher diversity parameter  $\alpha_{37} \approx 0.4$  versus the global  $\alpha_0 \approx 0$  influences predictions—is not merely statistical borrowing but a structural property of the underlying mixed Poisson architecture that enabled our model to assign nonzero predictions to county-disease pairs with zero training counts.

**4.2. Practical Extensions for Disease Surveillance.** The computational efficiency demonstrated in our California analysis, where both GG and GA models achieved stable convergence, suggests several immediate extensions:

**Multi-resolution modeling:** While we analyzed county-level data across 58 counties, the framework naturally extends to hierarchical structures like State → County → ZIP code, with each level maintaining its own subordinator process. The mixed truncated Poisson structure ensures tractability even as complexity grows.

**Dynamic surveillance:** Although our analysis treated years 2001–2014 independently for training, the framework’s Lévy-Itô foundation [4] enables continuous-time modeling where disease emergence follows time-evolving Poisson random measures, capturing seasonal patterns and emerging variants.

**Alternative disease spaces:** The beta-diversity patterns we observed arose from geographic proximity, but the framework can incorporate other distance metrics—genomic similarity between strains, human mobility networks, or environmental factors—by choosing appropriate Lévy measures  $\tau_j$  that encode domain-specific transmission dynamics [4, Section 7.5.3].

**4.3. From Local Patterns to Global Understanding.** Our analysis revealed clear geographic structure in disease diversity: southern counties exhibited higher alpha-diversity, while beta-diversity increased with geographic distance from reference counties. The precision of these estimates scaled with population size, demonstrating how the model appropriately weights information from data-rich versus data-sparse regions.

These patterns emerge from the four-component system  $(I_j, A_j, F_{j,\ell}, Z_j)_{j \in [J]}$  where:

- The coarse process  $Z$  captures the observed disease counts (our 36 rare diseases)
- The fine process  $I$  represents unobserved transmission chains
- The allocation process  $A$  determines which diseases manifest in each county
- The fragmentation operators  $F_{j,\ell}$  describe how statewide patterns decompose locally

This architecture enabled the key finding that even county-disease pairs with zero training counts could be predicted with meaningful uncertainty quantification—a capability essential for operational surveillance where the question “where will it appear next?” is paramount.

**4.4. Implications for Real-Time Implementation.** The framework’s robustness with sparse data—demonstrated by accurate predictions for diseases never observed in training—combined with the exact sampling procedures, enables:

- **Nowcasting:** Real-time estimation of current outbreak intensity from incomplete reporting
- **Forecasting:** Principled prediction to currently unaffected counties via the allocation machinery
- **Uncertainty quantification:** Full posterior distributions for resource allocation decisions

The same architecture that allows borrowing of statistical strength across California’s 58 counties could be applied to multi-site clinical trials, environmental monitoring networks, or financial contagion modeling—each instantiating the same fundamental duality with domain-appropriate specifications [4, Section 7.5].

These capabilities, grounded in the deeper mathematical architecture of simultaneous structural duality, point toward a unified framework for understanding how epidemiological information propagates through hierarchical systems—whether predicting the next county to report Listeriosis or identifying emerging transmission patterns—where the challenge is not just handling zeros, but understanding what they represent.

**5. Conclusion.** This work demonstrated the Poisson Hierarchical Indian Buffet Process (PHIBP) framework’s effectiveness in modeling rare infectious diseases across California’s 58 counties from 2001-2019. By analyzing 36 diseases with extreme sparsity of county-disease-year combinations, we showed how principled hierarchical modeling can extract meaningful patterns where traditional methods fail.

Our experimental results compared two specific PHIBP variants from the broader family of possible specifications: the gamma-gamma (GG) and gamma-alpha (GA) models. Among these two choices, GG consistently outperformed GA in test log-likelihood, demonstrating superior ability to capture rare disease patterns. Both models successfully predicted disease occurrences for county-disease pairs never observed in training data—a critical capability for surveillance systems where anticipating disease emergence in new locations is paramount. These GG and GA specifications represent only a small subset of the rich family of Lévy measures available within the PHIBP framework. The h-biased framework presented in Section 7 of [4] offers even greater flexibility, allowing for asymmetric allocation patterns and size-biased sampling that could better capture disease-specific transmission characteristics or preferential attachment dynamics in epidemic spread.

Geographic structure emerged clearly in our diversity analyses: beta-diversity increased with geographic distance between counties, while alpha-diversity showed regional clustering with southern counties exhibiting richer disease profiles. The precision of these estimates scaled appropriately with population size, with densely populated counties like Los Angeles showing the highest posterior certainty. These patterns, while robust under both GG and GA specifications, illustrate how different choices of subordinator processes can emphasize different aspects of the hierarchical structure.

These empirical successes reflect the model’s dual capacity to fragment global disease pools into county-specific patterns while simultaneously borrowing strength across locations to inform predictions. San Francisco’s elevated diversity parameter ( $\alpha_{37} \approx 0.4$ ) compared to the global baseline ( $\alpha_0 \approx 0$ ) exemplifies how the framework captures local heterogeneity while maintaining computational tractability through the mixed truncated Poisson representation.

The PHIBP framework’s success on this challenging infectious disease dataset demonstrates that its underlying mathematical architecture—grounded in coagulation-fragmentation duality—provides a principled approach to information sharing in hierarchical systems. The h-biased extensions could prove particularly valuable for modeling superspread events or hub counties that disproportionately influence disease transmission networks. Future applications might explore multi-resolution disease surveillance, continuous-time outbreak modeling, and incorporation of alternative distance metrics based on genomic similarity or mobility networks, potentially leveraging the h-biased framework’s ability to model preferential allocation patterns. More broadly, this unified framework for understanding how information propagates through complex hierarchical systems extends beyond epidemiology to microbial communities, clinical trials, and

other structured populations where the challenge is not just handling zeros, but understanding what they mean.

**Funding.** This work was supported in part by RGC-ECS grant 27304424 and RGC-GRF grants 16301521, 17306925 of the HKSAR.

## REFERENCES

- [1] BROOKS, S. P. and GELMAN, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7** 434–455.
- [2] FISHER, R. A., CORBET, S. A. and WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* 42–58.
- [3] GELMAN, A. and RUBIN, B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7** 457–511.
- [4] JAMES, L. F. (2025). Coagulation-Fragmentation Duality of Infinitely Exchangeable Partitions from Coupled Mixed Poisson Species Sampling Models. <https://doi.org/10.48550/arxiv.org/abs/2508.18668>
- [5] JAMES, L. F., LEE, J. and PANDEY, A. (2025). Poisson Hierarchical Indian Buffet Processes-With Indications for Microbiome Species Sampling Models. <https://doi.org/10.48550/arXiv.2502.01919>
- [6] JONES, P. C. T., MOLLISON, J. E. and QUENOUILLE, M. H. (1948). A technique for the quantitative estimation of soil micro-organisms. *Journal of General Microbiology* **2** 54–69.
- [7] LEE, M. D., WALWORTH, N. G., SYLVAN, J. B., EDWARDS, K. J. and ORCUTT, B. N. (2015). Microbial communities on seafloor basalts at Dorado Outcrop reflect level of alteration and highlight global lithic clades. *Frontiers in Microbiology* **6** 1470.
- [8] OF PUBLIC HEALTH, C. D. (2024). Infectious Diseases by Disease, County, Year, and Sex. <https://data.chhs.ca.gov/dataset/infectious-disease>. California Health and Human Services Agency Open Data. Updated June 7, 2025.
- [9] RICOTTA, C., SZEIDL, L. and PAVOINE, S. (2021). Towards a unifying framework for diversity and dissimilarity coefficients. *Ecological Indicators* **129** 107971. <https://doi.org/10.1016/j.ecolind.2021.107971>
- [10] WILLIS, A. D. and MARTIN, B. D. (2022). Estimating diversity in networked ecological communities. *Biostatistics* **23** 207–222.