Published in final edited form as:

Spat Stat. 2019 March ; 29: 243–267. doi:10.1016/j.spasta.2018.12.003.

Using a spatial point process framework to characterize lung computed tomography scans

Brian E. Vestal^{a,b,*}, Nichole E. Carlson^b, Raúl San José Estépar^c, Tasha Fingerlin^a, Debashis Ghosh^b, Katerina Kechris^b, David Lynch^d

^aCenter for Genes, Environment and Health, National Jewish Health, 1400 Jackson St, Denver, CO 80206, USA

^bDepartment of Biostatistics and Informatics, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO, USA

^cApplied Chest Imaging Laboratory (ACIL), Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

^dDepartment of Radiology, National Jewish Health, Denver, CO, USA

Abstract

Pulmonary emphysema is a destructive disease of the lungs that is currently diagnosed via visual assessment of lung Computed Tomography (CT) scans by a radiologist. Visual assessment can have poor inter-rater reliability, is time consuming, and requires access to trained assessors. Quantitative methods that reliably summarize the biologically relevant characteristics of an image are needed to improve the way lung diseases are characterized. The goal of this work was to show how spatial point process models can be used to create a set of radiologically derived quantitative lung biomarkers of emphysema. We formalized a general framework for applying spatial point processes to lung CT scans, and developed a Shot Noise Cox Process to quantify how radiologically based emphysematous tissue clusters into larger structures. Bayesian estimation of model parameters was done using spatial Birth–Death MCMC (BD-MCMC). In simulations, we showed the BD-MCMC estimation algorithm is able to accurately recover model parameters. In an application to real lung CT scans from the COPDGene cohort, we showed variability in the clustering characteristics of emphysematous tissue across disease subtypes that were based on visual assessments of the CT scans.

Keywords

Cox cluster process; Birth–death MCMC; Emphysema; COPD

1. Introduction

Chest Computed Tomography (CT) scans are a primary diagnostic and monitoring tool for many lung diseases. Typically, characterization of the CT image is done via visual

*Corresponding author at: Center for Genes, Environment and Health, National Jewish Health, 1400 Jackson St, Denver, CO 80206, USA. vestalb@njhealth.org (B.E. Vestal).

assessment, which can have poor inter-rater reliability, is time consuming, and requires access to trained assessors (Barr et al., 2012; Cavigli et al., 2009). Quantitative methods that reliably summarize the biologically relevant characteristics of an image are needed to improve the way lung diseases are characterized. The goal of this work is show how spatial point process models can be used to create a set of radiologically derived quantitative lung biomarkers of emphysema, one of the two major disease processes in Chronic Obstructive Pulmonary Disorder (COPD). We develop the framework generally, but show in particular how a Cox cluster process can characterize various emphysema patterns. The result of this model is a set of quantitative biomarkers that may then be used in a second stage of analysis to investigate associations with clinical, environmental, or genetic/genomic characteristics. Specifically, the Cox cluster process provides the number of emphysematous regions, along with both the average size and variation in size of the regions. In addition, we quantify the compactness of diseased tissue in these regions. We compute these measures by 2D slice from the top to the bottom of the lung which allows for a spatially varying investigation of the nature of diseased tissue.

1.1. Motivating data

Our work is motivated by the COPDGene study, a multicenter observational study designed to identify genetic factors associated with COPD (Regan et al., 2011). This cohort consists of approximately 10,000 smokers with and without diagnosed COPD across the GOLD stages (categorical staging system to represent severity of COPD). For each individual a series of inspiratory and expiratory chest CT scans were acquired. We focus our analysis on the inspiratory scans as these are primarily used for emphysema diagnosis and visual assessment in practice.

In general a CT scan can be thought of as a 3D reconstruction of an object using X-rays taken from various angles to generate slices that are then combined using computer processing. The underlying structure of the high resolution CT scans used here is a collection of points on a regularly spaced grid in a 3D space. Each of these points, which are termed voxels, is assigned a value in terms of Hounsfield Units (HU) that represents the density of the material at that voxel. The HU value, which measures the radiodensity of a substance with respect to that of distilled water, is then plotted as a Gray Scale (GS) image where darkness represents an absence of X-ray attenuation and lightness the opposite (e.g. air holes associated with emphysema appear as clusters of darkness while bone would be very bright). HU in CT scans range from -1024 to $1000+$ where normal lung tissue is around -400 HU and air is approximately -1000 HU. In COPD research, emphysematous tissue is often identified by a simple threshold mask (Lynch and Al-Qaisi, 2013). A voxel that is < -950 HU is classified as a low attenuation area (LAA) and indicative of emphysematous tissue (Fig. 1). The voxels labeled as LAA are the data we propose to model using a spatial point process framework. We note that other diseased tissue classifiers (e.g. Mendoza et al. (2012) or Hame et al. (2014)) could also be accommodated making the model generalizable to (a) new classifiers of emphysematous tissues and (b) other lung diseases (e.g. sarcoidosis).

1.2. Current approaches to quantifying emphysema

The most common quantitative summary of emphysema has been computing the percentage of LAA in the lung (Gu et al., 2015; Mohamed Hoesein et al., 2014; Nakagawa et al., 2016; Nambu et al., 2016; Matin et al., 2016). %LAA is associated with clinically relevant pulmonary characteristics such as FEV1 and FVC (Lynch and Al-Qaisi, 2013; Schroeder et al., 2013). Only weak genetic and biomarker associations with %LAA have also been found (Carolan et al., 2014; Kong et al., 2011). We posit that this could be in part due to the overly simplistic nature of %LAA as a measure since it is likely combining spatially heterogeneous disease patterns that might not show similar functional characteristics. For example, one might expect that 10% emphysema evenly spread throughout a lung could be quite different biologically from 10% emphysema where all of the diseased tissue is concentrated in a single lobe. Moreover, Kirby et al. found that visual assessment scoring provided independent information from the quantitative measures they investigated (Kirby et al., 2017).

Identifying clusters of diseased tissue as contiguous regions of LAA (Low Attenuation Clusters) has also been proposed (Lynch and Al-Qaisi, 2013; Pike et al., 2015). A major difference in our method is we relax the criteria for something to be labeled a cluster by not requiring them to be contiguous. This perhaps better reflects the stochastic nature of the image. The point process approach does rely upon having a definition of a diseased voxel. A significant amount of the work related to quantifying emphysema in CT scans has been focused on developing methodologies for better segmentation of emphysematous tissue in the lungs (Mendoza et al., 2012; Castaldi et al., 2014; Hame et al., 2014; Sluimer et al., 2006; Vegas-Sanchez-Ferrero et al., 2016). Recent work by Vegas showed that a test for emphysema in lung CT scans using a mixture of gamma distributions for the HU values converged to using a density mask with a -951 HU threshold (Vegas-Sanchez-Ferrero et al., 2016).

The remainder of this work is organized as follows: In Section 2 we lay out the general framework for applying spatial point processes to lung CT scans and formalize the cluster process to be applied in this instance. For parameter estimation we utilized a Bayesian framework and Spatial Birth–Death MCMC (BD-MCMC) as outlined by Moller and Stephens (Moller and Waagepetersen, 2003; Stephens, 2000). The BD-MCMC algorithm, along with some modifications to ease the computational burden due in part to size of the data sets (a single CT scan masked for only lung tissue can have on the order of 40 million voxels), will also be described. In Section 3 we examine simulation results along with an application to the motivating data example using CT scans from the COPDGene cohort study. Conclusions and future directions are explored in Section 4.

2. Methodology

2.1. A general spatial point processes framework for diseased tissue in lung CT scans

Consider a region of lung tissue $S \subseteq \mathbb{R}^d$ with Lebesgue measure $\mathcal{L}(S)$. In our examples we will fix S to be axial 2D slices of lung tissue, but the models easily generalize to 3D regions. In general, a spatial point process $\mathcal{X}$ is a random countable subset of the space S . For this

application $\mathcal{X}$ is the set of locations (voxels) in a lung CT slice that are defined as emphysema using a -950 HU density mask. In particular, we consider $\mathcal{X}$ to be a Poisson point process, which are driven by an intensity function $\lambda(\xi)$ where $\xi \in \mathcal{S}$. Conceptually the intensity function gives the “likelihood” of any particular point (voxel) being included in a realized set of locations generated by $\mathcal{X}$. Additionally, the number of points generated by $\mathcal{X}$ in a region $B \subseteq \mathcal{S}$ follows a Poisson distribution with mean defined by the intensity measure $\Lambda(B) = \int_B \lambda(\xi) d\xi$. A useful characteristic of lung CT imaging for emphysema is that we observe the entire space $\mathcal{S}$ on which the process can exist (i.e. $B = \mathcal{S}$), and we will utilize this to ease the computational burden for our proposed method. If the intensity function $\lambda(\xi) = \gamma$, i.e. it is constant for all ξ , then the point process is said to be homogeneous. If the intensity function is inhomogeneous, this can lead to clustering of points. Additionally, if $\lambda(\xi)$ is itself allowed to be stochastic (i.e. $\Lambda(B)$ is a random measure) then we obtain a class of Poisson point processes known as Cox processes. For example, a commonly used class of Cox processes is the log-Gaussian Cox process in which $\log(\lambda(\xi))$ follows a Gaussian process.

In this research we propose the use of another well known class of Cox processes to quantify emphysema in lung CT scans, the Shot Noise Cox Process (SNCP). In general, this is a hierarchical model where one (latent) point process generates cluster centers, then conditional on the cluster centers, and possibly a set of associated parameters or “marks” as well, a set of daughter point processes generate the observed points. This gives a point process $\mathcal{X}$ driven by random intensity function $\lambda(\xi) = \sum_{(c, \alpha) \in \Phi} \alpha k(c, \xi; \theta)$ where $k(\cdot, \cdot; \theta)$ is a kernel such that $k(c, \cdot)$ is a density for all $c \in \mathcal{S}$, which depends on parameters θ , and Φ is a marked Poisson process on $\mathbb{R}^d \times (0, \infty)$ with intensity function ζ . We can view $\mathcal{X} | \Phi$ as the superposition of independent Poisson processes $X_{(c_i, \alpha_i)}$ with intensity functions $\alpha_i k(c_i, \xi; \theta_i)$. As such, we can think of Φ as a center process generating a set of cluster centers $C = \{c_i, c_i \in B\}$ with marks α_i and θ_i , and the individual $X_{(c_i, \alpha_i)}$ as a cluster process with center c_i , intensity (or approximate expected size) α_i , and dispersion density $k(c_i, \xi; \theta_i)$. Then, with the specification of appropriate densities/priors for α and θ , $p(\alpha)$ and $p(\theta)$ respectively, the joint density of the data, cluster centers and parameters then factorizes as:

$$f(X, C, \alpha, \theta) \propto f_{\mathcal{X}}(X | C, \alpha, \theta) \cdot f(C) \cdot p(\alpha) \cdot p(\theta), \quad (1)$$

where first term is the density of a Poisson process, the second term is the density of the cluster centers C (i.e. another point process, typically a Poisson as well), and the final terms represent priors of other parameters in the intensity function.

2.2. The emphysema model

We propose a hierarchical Cox cluster process (the SNCP) to describe the clustering of diseased tissue that is clinically interesting for characterizing CT scans showing radiological evidence of emphysema. In this two level model, Level 1 is the observed spatial locations of diseased tissue based on applying the density masking technique with a -950 HU threshold. Next, we are interested in how individual diseased voxels cluster regionally (Level 2). The parametric model described below will allow for characterization of the CT slices with

respect to the degree of clustering, average cluster size, and the density/compactness of individual clusters in terms of the spread of points around the cluster center. Physical interpretations of each parameter are listed in Table 1.

2.2.1. Level 1: Diseased voxels—Let X be a realization of low attenuation areas (voxels) from point process $\mathcal{X}$ and C be a realized set of latent cluster locations from a point process $\mathcal{C}$. We model the locations of low attenuation pixels as an inhomogeneous Cox Process $\mathcal{X}$ on $\mathcal{S}$, where $\mathcal{S}$ is a 2-D axial scan of lung tissue (Fig. 1). For each $c_i \in C$ let $\mathcal{X}_i$ be a Cox process on $\mathcal{S}$ with random intensity function

$$\lambda_i^x(\xi | c_i, \alpha_i, \Sigma_i) = \frac{\alpha_i \cdot |\Sigma_i|^{-\frac{1}{2}}}{(2\pi)} e^{-\frac{1}{2}(\xi - c_i)^T \Sigma_i^{-1} (\xi - c_i)}. \quad (2)$$

This specifies that the low attenuation points scatter in a Gaussian shaped kernel around cluster center c_i with spread Σ_i and the average number of low attenuation voxels in cluster i , except at the boundaries, is α_i . Further, to account for low attenuation voxels that do not cluster, we also define a scatter process $\mathcal{X}_0$ as a homogeneous Poisson process with intensity function $\lambda_0^x(\xi) = \eta$ where $\eta \in \mathbb{R}^+$. This is likely to occur based on normal variation in lung CT scans based on factors such as inhalation level or BMI that could affect the distribution of HU values in a scan (Mishima et al, 1999). Invoking the superposition principle for Cox processes we can then define $\mathcal{X} = \left(\bigcup_{c_i \in C} \mathcal{X}_i \right) \cup \mathcal{X}_0$ where the random intensity function is $\lambda^x(\xi | C, \alpha, \Sigma) = \left(\sum_{c_i \in C} \lambda_i^x(\xi) \right) + \eta$. Now we define $X = \{x_i; x_i \in B\}$, the set of observed diseased tissue locations in $\mathcal{S}$, to be a realization of $\mathcal{X}$.

2.2.2. Level 2: Cluster centers—The cluster center locations are a marked homogeneous Poisson process $\mathcal{C}$ with associated intensity function $\lambda^c(\xi) = \gamma$ and intensity measure $\Lambda^c(S) = \gamma \cdot \mathcal{L}(S)$. Let $C = \{c_i\}$ be a realization of cluster locations from $\mathcal{C}$. Then each c_i has a pair of associated marks $\alpha_i \in \mathbb{R}^+$ and $\Sigma_i \in M_d^+$ (positive semi-definite matrices of dimension d). We define $\alpha = \{\alpha_i\}$ and $\Sigma = \{\Sigma_i\}$ and allow the individual α_i and Σ_i to be random effects generated from distributions $p_\alpha(\cdot)$ and $p_\Sigma(\cdot)$ that can also depend on additional hyper parameters. One such parameter we will utilize is μ_α , the average size of Level 2 clusters that will drive $p_\alpha(\cdot)$.

In the actual implementation we ultimately considered two different models for the cluster centers, a homogeneous Poisson process as described above and a non-homogeneous version. This was necessitated because in practice a single larger cluster is more likely to be divided into multiple clusters using the homogeneous Poisson. To lessen the chance of this happening, we considered a second model for the clusters that incorporates repulsion between the cluster centers. Specifically we used a soft-core model, which is an inhomogeneous Poisson process with intensity function $\lambda^c(\xi) = \gamma \prod_{i \neq j} P(c_i, c_j)$ where $P(c_i, c_j)$ defines some penalty (bounded between 0 and 1) for the pair of points c_i and c_j . There are

numerous options for the form of P , but we investigated a Strauss Process with $P(c_i, c_j) = \phi^{I_{r < \rho}}$ where r is the Euclidean distance between c_i and c_j , $\phi \in [0,1]$ is the penalty parameter, and $\rho > 0$ is a bandwidth parameter. Essentially this modification to the intensity function penalizes the likelihood for trying to place two cluster centers within the repulsive distance ρ , thereby only allowing for two or more centers to be closer than ρ if there is overwhelming data/evidence that they should be there. The resulting model fits are naturally sensitive to the choice of the repulsive distance, and we go into further detail on how we chose the specific values for our application in Section 2.3.2, and give general recommendations in the Discussion.

2.2.3. Other priors—We assumed the following priors for the remaining model parameters (all assumed to be mutually independent of each other and the cluster locations c_j): $\eta \sim \text{Gamma}(k, \lambda)$, $\log(\alpha_i) \sim N(\mu_\alpha, \sigma_\alpha^2)$, $\mu_\alpha \sim N(a, t^2)$, $\Sigma_j \sim \mathcal{FW}(U, v)$. Subsequently we can combine the results from both levels to write the joint distribution of all sets of locations and marks as

$$f(X, C, \alpha, \Sigma, \mu_\alpha) \propto \left(e^{-\Lambda^X} \prod_{x_i \in X} \lambda^x(x_i | C, \alpha, \Sigma) \right) \cdot \left(e^{-\Lambda^C} \prod_{c_i \in C} \lambda^c(c_i) \right) \cdot \left(\prod_{\alpha_i \in \alpha} p_\alpha(\alpha_i | \mu_\alpha) \right) \cdot \left(\prod_{\Sigma_i \in \Sigma} p_\Sigma(\Sigma_i) \right) \cdot p_{\mu_\alpha}(\mu_\alpha). \quad (3)$$

2.3. Estimation: Spatial birth–death MCMC

The point process model parameters are estimated using a spatial BD-MCMC algorithm (Moller and Waagepetersen, 2003). This method is appealing in this scenario because it allows the number of mixture components (i.e. clusters/air holes) to be unknown, and this quantity is then estimated along with the other parameters. The BD-MCMC algorithm alternates between two states at each iterations: the BD process, and the traditional MCMC step. In the BD portions, model components (clusters) are systematically added and removed based on their contribution to the likelihood. Once this is finished, a set of cluster centers is obtained and then the remaining parameters are updated using standard MCMC techniques (e.g. Gibbs sampling). The overall BD-MCMC algorithm iterates between these two states, and ultimately what is obtained is a posterior sample of model parameters that takes into account the uncertainty associated with not knowing the number of mixture components. Details of the BD-MCMC algorithm and calculations of relevant quantities can be found in Appendix A. The implementation, available in the *sncp* R package, is done as a hybrid of R and C++ using the *Rcpp* and *RcppArmadillo* packages (Eddelbuettel et al., 2011; Eddelbuettel and Sanderson, 2014).

2.3.1. Computational solutions—Naive approaches to implementing a BD-MCMC algorithm are computationally prohibitive for CT images given the size of the data and the complexity of the SNCP likelihood calculations. Here we outline computational solutions that make fitting point process models for this problem more efficient. First, we developed an efficient birth surface for introducing new cluster centers into the model to speed mixing

and convergence. Second we show how leveraging the fact that the entire space upon which the process can exist is observed (i.e. emphysema can only exist within the lung which is considered to be observed in its entirety) allows us to simplify some integral calculations.

Let $b(C, \xi) \cdot p(\theta')$ be the birth rate and $d(C \setminus \xi, \xi)$ be the death rate for cluster centers C , where ξ is a new center location and θ' are new cluster specific marks associated with ξ drawn from the prior $p(\theta')$. From these we obtain an overall birth rate $\beta(C) = \int_S b(C, \xi) d\xi$ and overall death rate $\delta(C) = \sum_{\xi \in C} d(C \setminus \xi, \xi)$ that determine the event rates in the BD process.

The birth and death rates are related via detailed balance to ensure reversibility of the chain:

$$h(X | C, \theta) p(C) p(\theta) b(C, \xi) p(\theta') = h(X | C \cup \xi, \theta \cup \theta') p(C \cup \xi) p(\theta \cup \theta') d(C \setminus \xi, \xi), \quad (4)$$

where $h(X|C)$ is the unnormalized density for the observed point process X , $p(C)$ is the point process density for the latent cluster centers C , and $p(\theta)$ is the prior for additional model parameters. To simulate the BD process we can define the birth rate b a priori and then solve for the death rates d at each iteration, or vice versa. The latter option is often computationally expensive and/or analytically intractable (as is the case for our proposed model), so a common setup is to fix the overall birthrate β at 1 (i.e. $b(C, \xi) = 1 / \mathcal{L}(S)$) and then propose new mixture components as draws from a uniform distribution on S . This leads to relatively straightforward computation of death rates, but is inefficient for the BD process since proposing new mixture components uniformly over S will likely lead to many “bad” proposals in areas where no diseased tissues were observed. These components are then immediately removed, and this behavior can lead to slow mixing and convergence.

To alleviate this issue we have devised a novel method for defining the proposal distribution: the observed point pattern X is passed through a non-parametric smoother and then normalized so that it still integrates to 1 (i.e. overall birth rate $\beta(C) = 1$). That is, we create $b(C, \xi)$ as

$$b(C, \xi) = \frac{\sum_{x_i \in X} \phi(\xi | x_i, W)}{\int_S \sum_{x_i \in X} \phi(\xi | x_i, W) d\xi}, \quad (5)$$

where $\phi(\xi | x_i, W)$ is the density of a multivariate Gaussian distribution centered at x_i with covariance matrix W . This gives a proposal distribution that is weighted more towards the regions where data was observed, and intuitively gives the BD process a better chance at proposing mixture components that will remain in the model. Note that this surface remains static throughout the BD-MCMC and does not depend on the current configuration of C . Therefore, we do not run into the same computational burden of needing to re-define the birth rate at each iteration as one would need to do if the death rates were fixed instead. In practice we used Gaussian kernels to construct the smoothed surface by placing a multivariate normal kernel on each diseased pixel with a covariance matrix $W = 100 * I_2$. This has a similar motivation to a birth rate proposed by Lawson, but differs significantly in that it does not depend on the current configuration of C , and thus does not need to be recomputed at each iteration (Lawson and Denison, 2002).

A second computational modification was to assume that the integral in the normalizing constant for the observed points $\mathcal{X}$ is equal to 1. The intensity function here is of the form $\lambda^{\mathcal{X}}(\xi) = \sum_{c_i \in C} \alpha_i k(\xi | c_i, \theta_i)$ where k is the dispersion density for a cluster centered at c_i . The intensity measure (that goes into the normalizing constant for $f_{\mathcal{X}}(\mathcal{X} | C, \mathbf{a}, \boldsymbol{\theta})$) is then $\Lambda^{\mathcal{X}} = \int_S \lambda^{\mathcal{X}}(\xi) d\xi$, and computation of this involves numerically approximating integrals of these densities over an irregular space S (outline of the lung). This is very expensive computationally and slows the BD-MCMC process immensely. By assuming that all the integrals are always equal to 1, run times were greatly reduced but the interpretations of fitted parameter estimates are slightly different. To see this, consider the intensity measure for an arbitrary cluster c_i :

$$\Lambda_i^{\mathcal{X}} = \int_S \lambda_i^{\mathcal{X}}(\xi) d\xi = \alpha_i \int_S k_i(\xi | c_i, \theta_i) d\xi. \quad (6)$$

When c_i is near the interior of S , $\int_S k(\xi | c_i, \theta_i) d\xi \approx 1$ and $\Lambda_i^{\mathcal{X}} \approx \alpha_i$. However, when c_i is near the boundary of S , $\int_S k(\xi | c_i, \theta_i) d\xi < 1$ and thus $\Lambda_i^{\mathcal{X}} < \alpha_i$. By assuming that $\int_S k(\xi | c_i, \theta_i) d\xi = 1$ for all c_i we end up estimating Λ_i instead of α_i . In practical interest, the average size of the cluster (Λ_i) is likely more useful than the latent parameter α_i . Moreover, the entire space on which the process $\mathcal{X}$ can exist is observed so there is no fear of underestimating the sizes of clusters because of artificial boundaries due to a window of observation since emphysema cannot occur outside of the lungs. While our main concern with this approximation was with respect to α_i , this also has the potential to affect the estimates of c_i and θ_i . However, in simulations to investigate the behavior of this modification we found that estimates still closely tracked the quantities of interest (results not presented) and model fits were not compromised.

2.3.2. Estimation implementation—The BD-MCMC algorithm was run using the following values for the general priors listed above (chosen based on clinical input from investigators): $\log(\alpha_j) \sim N(\mu_a, 2)$, $\mu_a \sim N(4, 6)$, $\Sigma_j \sim \mathcal{IW}(40 \cdot I_2, 5)$, and $\eta \sim \text{Gamma}(0.01, 0.01)$. For the center process $\mathcal{C}$ we assumed a Strauss Process prior with $\gamma = 4 / \mathcal{L}(S)$, $\phi = 10^{-9}$, and $\rho = 20$. These values were chosen to make sure that the issues observed regarding breaking of clusters into multiple components were virtually negated. Note that estimates of the clustering are sensitive to the value of ρ and can be directed to be more local (small values of ρ) or more regional with large values of ρ (Fig. B.6 in Appendix B). The value of 20 was chosen based on visual inspection of model fits on several example scans as it provided clustering on the scale that was deemed relevant for this application. Posterior chains of length 50,000 were generated and summarized discarding nothing as burn-in and no thinning. For proposal distributions in the MCMC we used a random walk for η , $\log(\alpha_j)$, and Σ_j . Specifically, $\eta^{(t+1)} \sim U(\eta^{(t)} - .01, \eta^{(t)} + .01)$ (reflexive at 0), $\log(\alpha_j^{(t+1)}) \sim N(\log(\alpha_j^{(t)}), .5)$, and $\Sigma_j^{(t+1)} \sim \mathcal{IW}(\Sigma_j^{(t)}, 15)$; μ_a could be drawn from an appropriate Normal distribution due to conjugacy. The cluster center locations c_i were also updated with a random walk using a uniform square centered on the current value with sides of length 5. Mixing and convergence were assessed using visual inspection of trace plots for

μ_α and $|C|$ (since these parameters remain present throughout the BD-MCMC), as well as visual assessment of movies showing the evolution of the BD-MCMC chain (see Supplementary Materials for an example movie). The random walk step sizes were chosen for acceptance probabilities of roughly 25%–50%. Parameters were initialized using random draws from the appropriate prior. Posterior distributions were summarized using the median and 90% equal tailed credible intervals.

3. Results

3.1. Simulations

We conducted 2 broad simulations, one with moderately sized clusters (proof of concept) and a second using sets of parameters derived from model fits on the exploratory data set described below. All of the simulated data sets used the same templates space $\mathcal{S}$, an example axial slice from a lung CT scan. The cluster centers were generated from a Strauss Process ($\phi = 10^{-4}$, $\rho = 20$ or 10) using a spatial birth–death algorithm for simulation. Given a set of cluster centers, μ_α , α and Σ were simulated from the prior distributions. Conditional on these values, the number of points for cluster i was simulated as $N_i \sim \text{Poisson}(\Lambda_i^x)$ and the locations X_i generated by cluster i were simulated as $X_i \sim \mathcal{MVN}(c_i, \Sigma_i)$. See Appendix C for more details on the simulation process.

3.1.1. Simulation 1: Proof of concept—To examine the spatial BD-MCMC algorithm’s ability to identify low attenuation clusters (and the associated parameters describing them), 500 data sets were simulated using a set of priors that on average generated point patterns with moderately sized clusters. The parameters used to simulate the data are listed in Supplementary Table C.8 in Appendix C, along with the priors that were chosen for the BD-MCMC fits. Here the algorithm was run for 50,000 iterations (nothing discarded as burn-in), and performance was evaluated using bias, RMSE, and the coverage of 90% posterior credible intervals for $|C|$, μ_α , and η .

Fig. 2 (top row) shows the general progression of the simulations from the simulated point pattern to the posterior summaries of the BD-MCMC chain. Moving from left to right, the first plot shows the point pattern, while the second shows the posterior distribution for the locations of cluster centers. Note that in general the center locations are estimated with high spatial precision (i.e. small areas of red voxels that are consistently estimated to be cluster centers). The third and fourth panels show the posterior predictive distribution and the estimated clustering of points from an iteration of the BD-MCMC towards the end of the chain respectively. Additionally, Table 2 shows numerical summaries averaged over the 500 simulated data sets. For the number of clusters $|C|$ there is a trend towards slight overestimation. This was typically only a problem when there were excessively large (area wise) and densely packed clusters present in the simulated slices (Supplementary Fig. B.8 in Appendix B). The RMSE and bias for $|C|$ were both small and the 90% credible interval coverage was good at 87%.

We observe a slight underestimation of the average size of clusters on the log scale, μ_α . This is likely the result of two separate issues: the slight overestimation of the number of clusters

and the integral approximation. However, the effect was minimal on the RMSE. The credible intervals for μ_a were typically conservative (i.e. too wide) with the 90% interval coverage being roughly 99%. The results for η , the random scatter parameter, were largely the same as for $|C|$ with a small bias and RMSE combined with good credible interval coverage at 88%.

3.1.2. Simulation 2: Real data parameters—A second set of simulations were derived using the model parameters that were obtained from applying the BD-MCMC algorithm to real CT slices as described in the following section. Here 6 combinations of model parameters were chosen, and from each combination 100 2D point patterns were simulated. The 6 parameter combinations were chosen as the complete sets from the 2D slices that had the min, median, or max estimated $|C|$ or μ_a from the real slices (Supplementary Table C.9 in Appendix C). An additional modification was a shortening of the bandwidth parameter ρ from 20 to 10 in the Strauss Process used to simulate cluster centers to explore what happens when this quantity is misspecified in the BD-MCMC algorithm where $\rho = 20$ was used.

The numerical summary of Simulation 2 can be found in Table 2. Additionally, Fig. 2 (bottom row) shows the progression from simulated data to graphical summaries of the BD-MCMC chain. The major difference here is we observed a larger bias and RMSE for $|C|$, and lower coverage of the 90% credible intervals. This underestimation is almost certainly a result of using the smaller repulsion radius for generating the simulated data sets than what was assumed in estimation. However, this illustrates how the choices of various parameters can control the level at which the point process model looks for cluster components. While the priors put on the various parameters and hyper-parameters in the SNCP did not seem to have much effect on the posterior parameter estimates, the choice of penalization parameters can have a larger impact. By making the radius too small, the flexibility of the model allows even moderately sized clusters to be broken up into multiple components to accommodate small divergences from elliptical clusters implied by the multivariate normal dispersion densities assumed. However, going too far in the other direction implies that the model is searching for more regional clusters that are likely to be at a spatial resolution that is lower than we are interested in for this application. Thus, using expert clinical input is important to focus the model on an appropriate scale for the given problem.

3.2. Real data application: CT scans from the COPDGene cohort

We applied the proposed methods to two separate data sets pulled from the COPDGene baseline cohort: an exploratory set of 13 individuals randomly sampled from subjects with mild to moderate quantitative emphysema (%LAA), and a second set with 200 subjects sampled to investigate how model parameters compared between and within 10 emphysema subtypes.

3.2.1. Exploratory set—The proposed model was initially applied to inspiratory CT scans (standard reconstruction) from 13 individuals randomly selected from the COPDGene cohort that had between 5%–15% LAA (considered mild to moderately diseased). These were used to explore characteristics of the BD-MCMC algorithm on real data and to design

a simulation based on fitted model parameters from these scans. In each individual, each lung was segmented from surrounding tissue and then divided into thirds by volume axially. Within each third the axial slices corresponding to the 25th, 50th, and 75th percentiles by volume were analyzed to give a representation of patterns throughout the lung. Observed point patterns were created by applying the -950 HU density mask, and both lungs within a subject were used in this exploratory set giving a total of 234 slices examined. The parameters that we focused on include $|C|$, μ_a , η , and the average area of a cluster as determined by the area of the 75th percentile ellipsoid based on the Σ_i . The values presented in this section and the next represent normalized values rather than the raw estimates from each 2D slice. That is, parameters were normalized by slice area (e.g. we normalize $|C|$ as $|C| / \mathcal{L}(S)$ for each CT slice), and values are presented as rates per 20,000 pixels. This was done because each slice contains a varying number of pixels, as well as the fact that individuals are not all scanned at the exact same resolution (i.e. pixel spacing).

The results from the 234 real slices are summarized in numerical form in Table 3 and with histograms showing the distributions of point estimates using posterior medians in Supplementary Fig. B.7 in Appendix B. Fig. 3 shows a graphical progression from observed data to several visual summaries of the model fit for two example slices. In this example we see what appear to be visually distinct clustering patterns between 2 slices that have roughly the same %LAA. For the slice on the top row, estimates are $|C| = 18.79$, $\mu_a = 3.575$, $\eta = 0.0002$, and average area = 121.35. Alternatively, the slice on the bottom row has parameters $|C| = 25.04$, $\mu_a = 3.387$, $\eta = 0.0079$, and average area = 197.10. While the average number of pixels per cluster is roughly the same, the slice on the bottom has about 33% more clusters which on average cover about 50% more area. The estimate of the scatter process is also about 30 times as large. This seems to suggest less densely packed clusters and significantly more random scatter, which could be evidence of more panlobular emphysema in the lower scan versus centrilobular in the upper one.

3.2.2. Analysis set—A second, larger data set was used for the broader application of the proposed model. Here, 200 individuals were selected from the COPDGene cohort based on visual assessment subtypes listed in Table 4 (see Appendix D for more details). 20 subjects were randomly selected from each of the 10 subtypes, and then their standard reconstruction inspiratory CT scans were segmented, sampled, and density masked as described for the exploratory set. One lung was randomly selected for analysis in each individual, giving 1800 total slices.

Table 5 has a numerical summary of the model parameters broken down by subtype. Here model parameters were collapsed to within subject means across the 9 slices analyzed in each. Table 6 shows the rank-correlations between $|C|$, μ_a , η , and the average cluster size. Interestingly the size and number of clusters are positively correlated where the slices with the largest average size also tended to have the most clusters present per 20,000 pixels.

Fig. 4 shows boxplots for $|C|$, μ_a , η , and the average area of a cluster, again broken down by subtype. Here we see that some subtypes show more variation in model parameters than others, but there also appears to be significant between-subtype variation. Indeed, simple ANOVA models (using subject level means) for each parameter across the subtypes are all

highly significant ($p = 0.0001$). Note that the for $|C|$, μ_a , and the average cluster area the patterns are quite similar. For η (measure of random scatter, or amount of pixels that do not exhibit clustering) we see relative uniformity in 9 of the subtypes and a large increase in the other. The subtype that has the significant increase here is a discordant group that were not found to have significant emphysema based on visual assessment, but had $> 10\%$ LAA. This result could give some insight into this inconsistency as a large amount of random scatter in the SNCP would be indicative of diffuse disease without obvious structure that would be difficult to identify visually.

In the three subtypes that showed little to no visual evidence of emphysema (subtypes 1–3), the summary measures are all on the small side and quite similar across the groups. In terms of $|C|$ there are some subjects in these no emphysema groups that have values overlapping with what is seen in the subtypes characterized by more visual evidence of emphysema. However, a larger distinction is seen for average size (μ_a) where the expected number of pixels per cluster is quite small (≈ 20). For random scatter, the estimated values of η are relatively small and are consistent across these 3 groups. While the average area is roughly the same for subtype 3 as compared to 1 and 2, there is more variation and both the 75th percentile and maximum values are notably higher.

CT subtype 4 (mild emphysema) is where a transition starts to occur in that we see larger values on average for all of the model parameters. This subtype also shows the largest variability in most model parameters (the exception being η) where estimated values overlap what was seen in subjects that showed no signs of emphysema and those that had moderate to severe emphysema. While these lungs tended to have both more clusters and slightly more pixels per cluster (≈ 33) than what is seen in subtypes 1–3, the distribution of area of clusters is closer to subtype 3 than it is to the more severe subtypes. There is also an increase in the average amount of scatter observed with an almost doubling of η as compared to subtypes 1–3.

The next grouping of subtypes based on similar model parameters is 5–8. Three of these groups (5–7) are characterized by moderate to severe emphysema and are distinguished by where in the lung the predominant amount of disease exists (upper lobe, lower lobe, diffuse). The upper lobe predominant diseased lungs seem to have consistently higher values for all 4 parameters as compared to the lower lobe predominant diseased subjects, but the differences are small. The diffuse disease group (subtype 7) had the largest average $|C|$, μ_a , and average area of any group. The paraseptal emphysema group (subtype 8) looks remarkably similar to these other moderate to severe subtypes. This could suggest that paraseptal emphysema only co-occurs with mild or severe centrilobular or panlobular emphysema.

The last two groups (subtypes 9 and 10) are discordant groups where visual and quantitative emphysema (i.e. %LAA) do not agree. This bears out in our model parameters as well with subtype 9 (visual emphysema without quantitative emphysema) falling in between subtypes 1–3 (normal or no emphysema) and subtype 4 (mild emphysema). Subtype 10 (quantitative without visual emphysema) appears similar to subtypes 5–8 for values of $|C|$, μ_a , and average area. However, there is a large change in the random scatter process η in these subjects where the mean value in this group is 5–20 times larger than any other subtype.

This could explain why the visual assessment is so different from the quantitative results since there is a lot of low attenuation pixels that are not showing evidence of clustering, and this lack of structure could be leading to a visual assessment of no emphysema. Another explanation is that the CT scan values for these scans could have been affected by outside issues causing a general shift in the HU values to be more negative (darker on visualization), and this is what is being picked up on in the estimates of η since there does seem to be evidence of some clustering in these scans as well. A downward shift in HU values could also obscure the visual assessment by masking the true clustering in the visualization of the CT scan.

Beyond looking at subject level means across the subtypes, we believe that investigating the within-subject variation in model parameters as the lung is traversed could also yield valuable insight into disease subtypes. For example, Fig. 5 shows the entire trajectory from the bottom of the lung to the top for μ_a (see Appendix B for $|C|$, η , and average cluster area) across the 9 slices taken within each subject and split up by CT subtype. We observe a wide variety of within-subject variation where some subjects have parameter estimates that are almost constant across all 9 slices analyzed while others show much larger variability. Additionally, some subtypes exhibit more unique patterns than others with a certain shape to the curves being replicated by multiple individuals within the subtype. This is most noticeable in groups 5 and 6, which is consistent with these subtypes being defined by a lobar-predominant disease pattern (lower and upper respectively).

4. Discussion

In this work we have proposed a new framework for quantifying radiologically defined emphysema in lung CT scans using spatial point processes to describe the occurrence of diseased voxels and how they cluster into larger structures. Model parameters that have a straightforward physical interpretation are estimated under a Bayesian paradigm with a spatial BD-MCMC algorithm. The simulations conducted suggest that the model can accurately identify quantities of interest, while applications to real CT scans show variability in both the simple summary measures and the trajectories investigated so far. Indeed, we did observe separation between classes of visual assessment subtypes in either mean behavior using within-subject means, or in using the entire trajectory through the lung of model parameters. Using just axial slices seemed to separate the subtypes into two general groups, mild to no disease and moderate to severe disease. A potential extension would be to use sagittal and coronal slices as well since variation over these planes could provide supplementary information as well. However, we did observe noticeable trends in the subtypes that are characterized by lobar predominant emphysema, so using an axial trace could quickly identify these cases.

Another extension could be to add an additional layer of hierarchy onto the model. In this 3 level model, level 1 would still be the observed disease tissue voxels, level 2 would still describe how the observed voxels cluster at a local level but would themselves be distributed as a SNCP rather than a Strauss process. Then at level 3, we could model the centers of regional clusters around which the smaller level two clusters are distributed. Indeed the results from the BD-MCMC seem to point that this may be more appropriate than the 2 level

model used since there is a rather large discrepancy in the scale in which the clusters exist. For example, in the same CT slice we can see both small, densely packed clusters being identified within the contour of larger, more diffuse, and more regional clusters. The major issue with this extension will be the additional computational burden it would create.

A previous study had a similar motivation and attempted to quantify the complexity of terminal airspace geometry (Mishima et al., 1999). Here they found that the cumulative size distribution of LACs followed a power law that was characterized by an exponent D , which they interpreted as a measure of airspace complexity. An interesting piece of this study was that COPD subjects that did not have abnormally high %LAA still had lower values of D than the healthy subjects, but D did not correlate well with typical lung function measures. Moreover, they concluded that as LACs coalesce into larger structures, terminal airspace complexity will decrease (as measured by D in their models) while leaving %LAA relatively steady. We believe that the measures derived from the point process model described here should show similar results in that for a given %LAA, we can extract further information about the disease process based on the clustering. A major difference in our method is we relax the criteria for something to be labeled a cluster by not requiring them to be contiguous. This could be beneficial when using the standard density masking technique as noise in the CT scan reconstruction can artificially break up clusters into non-contiguous pieces that our method could potentially reconstruct into a single component. Furthermore, the straightforward biological interpretation of parameters (e.g. number of clusters and average size of clusters) in our point process model as compared to the measure D from Mishima et al. could help adoption in clinical practice.

One potential limitation of the method described here is the reliance on the -950 HU density masking to obtain a set of diseased tissue locations. Intuitively the use of single threshold for all scans seems less than ideal given that the entire distribution of HU values can be affected by issues such as the subject's BMI or their ability to fully inflate their lungs for inspiratory scans (Mishima et al., 1999). Indeed, there are additional methods for quantifying emphysema using, for example, local histogram analysis or hidden Markov Models (Mendoza et al., 2012; Castaldi et al., 2014; Hame et al., 2014). However, these were not used here for two reasons: availability of software implementing these methods, and the prevalence of density masking in clinical research and practice. Additionally, a recent paper proposing a test for the presence of emphysema using a mixture of gamma distributions found that their method converged to using a density mask with a -951 HU threshold (Vegas-Sanchez-Ferrero et al., 2016). Therefore the -950 HU density mask seems appropriate while acknowledging that being able to apply better pixel classification techniques would also benefit the method proposed here.

A more general limitation to using the BD-MCMC algorithm developed here to fit the SNCP is the need for a repulsive process between the cluster centers. When this piece was dropped (i.e. the prior for $\mathcal{C}$ was a homogeneous Poisson process) we observed even moderately sized clusters were split apart into several components. While this allows for extremely flexible estimation of the underlying intensity function at Level 1 $\lambda^X(\xi)$, it destroys our interpretation of clusters representing unique local LAA structures. Using the Strauss process helped ameliorate this issue, but it requires choosing a distance at which the

repulsion becomes active. If this value is chosen to be too small then over-fitting will likely still occur (Fig. B.6 in Appendix B). If it is chosen to be too large then multiple local clusters will likely be collapsed into more regional structures. Consequently, having expert clinical opinion on the approximate scale of the clustering is crucial. We chose the value in our application based on both the size of the clusters we were expecting and also how close we would expect unique clusters to be separated by. This was done by fitting the SNCP to several example CT slices with a range of repulsion distances ρ for each, and then examining the model fits to determine which value gave clusters that were generally on the scale that were deemed to be physically and biologically relevant. That value was then fixed for the analysis of all CT scans in the study. While the value we used for the real data application should work well for studying emphysema in other sets of CT scans, if one were to apply this model to a different disease process (e.g. sarcoidosis or fibrosis), a similar process with visual assessment of some example model fits would be needed to confirm that the chosen values are performing as expected before setting that repulsive distance for the analysis of all scans.

A second limitation is the computational feasibility of fitting the point process models in a large scale clinical application. The current implementation, available as part of the *sncp* R package, takes several minutes to numerous hours for each 2D slice from the real CT scans. The median run time for the 1800 slices examined in the analysis data set was about 50 min, while the maximum was almost 2 days. The SNCP density is computationally expensive and the BD process requires evaluating this many times per MCMC iteration so the run times grow quickly depending on the number of diseased tissue locations and the number of estimated clusters. However, with access to the right computing environment where slices can be analyzed in parallel it can greatly reduce the time needed to generate summary measures or a trajectory for a given subjects lung(s). Additionally, without some intensive post processing of the BD-MCMC chains, we are not able to easily summarize individual clusters because of the loss of identifiability due to label switching. This could be a limitation in longitudinal studies and in clinical practice if physicians are not satisfied with just summaries of mean behavior, and instead require cluster-level information.

By using the SNCP in combination with the BD-MCMC estimation we have a unified framework that can both discover the number of emphysematous lesions and quantify specific characteristics about the spatial patterns they exhibit. While both the model and estimation procedure are relatively complex, the SNCP has been chosen in a way so that the parameters of interest have a straightforward physical interpretation. With the release of the *sncp* R package, which contains a vignette explaining how to use the software along with access to the simulated data sets presented in this work and the scripts to reproduce the results, the difficulty in accessing this type of model is also significantly decreased. In terms of clinical adoption, we can envision the estimation of the SNCP model parameters being incorporated into a CT processing pipeline that will automatically generate summaries for review by a radiologist any time a new scan is obtained. Access to the *sncp* R package should allow for model summaries to be generated within a few hours to a few days from when a scan is acquired. Along with estimates for the number of emphysematous clusters/lesions ($|C|$), the average size of these lesions (μ_a), the average amount of diffuse disease (η), and the trajectories of these parameters when going from the bottom of the lung to the

top can provide a quick summary of the spatial distribution of diseased tissue for review by a radiologist, or even act as a flag for which scans require further visual assessment in terms of emphysema. Ultimately, we believe that the results presented here give credence to the hypothesis that the proposed model and summaries will better leverage the rich spatial information available in the lung CT scans that is discarded in the current standard quantitative measure %LAA. We further hypothesize that the metrics derived from the new models proposed here, because of this retention of more spatial information, will be more useful in identifying genetic or biomarker signatures while still providing a significant dimension reduction from the original CT scans.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This work was supported by NHLBI R01 HL089897 and R01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.⁵

Appendix A.: Details of spatial birth–death process

See Table A.7.

Table A.7

Common notation.

Symbol	Meaning
$\mathcal{X}$	A spatial point process
$X = \{x_i\}$	A realization of point process $\mathcal{X}$
$ X $	The cardinality (size) of an observed point configuration X
B	The space upon which a point process $\mathcal{X}$ is defined
ξ	An arbitrary point in B
$\mathcal{L}(B)$	The Lebesgue measure of a space B

In the BD portion of the model fitting algorithm we focus on generating a spatial birth–death process conditional on the current estimates of all other model parameters. Note that we will use a general notation of $\mathcal{Y}$ and Y for a spatial BD process in this section, but one can think of this as being applied to $\mathcal{C}$ from the proposed model. The distribution of a spatial BD process $\mathcal{Y}_t$, $t \in \mathbb{R}^+$ can be characterized by a birth rate b and a death rate d which are non-negative functions defined on $N_{lf} \times B$ (here N_{lf} denotes the collection of all possible locally finite point configurations in B). We can then define, for some $Y \in N_{lf}$

$$\beta(Y) = \int_B b(Y, \xi) d\xi \quad (\text{A.1})$$

$$\delta(Y) = \sum_{\xi \in Y} d(Y \setminus \xi, \xi) \quad \text{if } Y \neq \emptyset \quad (\text{A.2})$$

with $\delta(\emptyset) = 0$. Here $\beta(Y)$ and $\delta(Y)$ can be thought of as overall birth and death rates respectively for some configuration Y . Using these two we can define

$$\tau(Y) = \beta(Y) + \delta(Y) \quad (\text{A.3})$$

and this quantity will describe the overall event rate for a Poisson process in virtual time t that will govern the adding and removing of components from the spatial BD point process $\mathcal{Y}_t$. Indeed, $\mathcal{Y}_t$, $t \geq 0$ is right-continuous and piecewise constant except at the jump times $T_1 < T_2 < \dots$. Given an initial state $Y_0 \in N_{lf}$ set $T_0 = 0$, $\mathcal{Y}_0 = Y_0$, and $J_m = \mathcal{Y}_{T_m}$ for $m = 0, 1, \dots$, and suppose we condition on $J_0, T_0, \dots, J_m, T_m$ where $\tau(J_m) > 0$, then

$$T_{m+1} - T_m \sim \text{Exp}(\tau(J_m)) \quad (\text{A.4})$$

Essentially we are defining the Poisson process for changes in $|\mathcal{Y}_t|$ in terms of the exponentially distributed inter-arrival times of events (addition or deletion of points from $\mathcal{Y}_t$). Further more, if we condition on T_{m+1} , then either a birth or death occurs at time T_{m+1} where

- A birth occurs with probability $\beta(J_m)/\tau(J_m)$, in which case $J_{m+1} = J_m \cup \xi_m$ where the newborn point comes from the density

$$\bar{b}(J_m, \xi) = \frac{b(J_m, \xi)}{\beta(J_m)} \quad (\text{A.5})$$

Any marks are drawn from their appropriate distributions as well.

- A death occurs with probability $\delta(J_m)/\tau(J_m)$, in which case $J_{m+1} = J_m \setminus \xi_m$ where the point to be deleted ξ_m comes from the density

$$\bar{d}(J_m, \xi) = \frac{d(J_m \setminus \xi, \xi)}{\delta(J_m)} \quad (\text{A.6})$$

Now all that remains is to specify b and d such that we maintain reversibility (i.e. we can move around the entire sample space without restriction such that it is possible to obtain any $Y \in N_{lf}$). If we let $h(Y)$ represent the unnormalized density for a process $\mathcal{Y}$ with respect to the unit Poisson, then for all $Y \in N_{lf}$ and $\xi \in B \setminus Y$ the relationship

$$h(Y)b(Y, \xi) = h(Y \cup \xi)d(Y \setminus \xi, \xi) \quad (\text{A.7})$$

provides a detailed balance condition. With the additional assumption that $E\beta(Y) < \infty$ with respect to $h(Y)$ then the spatial BD process constructed in this way is reversible with respect to $h(Y)$. From the detailed balance condition we are presented with two options for specifying the BD process: fix the birth rate b and solve for the death rate, or fix the death

rate d and solve for the birth rate. As described in the main manuscript, we derived a novel data-driven birth rate. Now that we have $b(Y, \xi)$ and $h(Y)$ specified by the model, the only thing that remains is to solve for the death rates $d(Y \setminus \xi, \xi)$ by solving

$$d(Y \setminus \xi, \xi) = \frac{h(Y)b(Y, \xi)}{h(Y \cup \xi)} \propto \frac{b(Y, \xi)}{\lambda^*(Y, \xi)} \quad (\text{A.8})$$

where $\lambda^*(Y, \xi)$ is known as the Papangelou conditional intensity. Here a heuristic interpretation of $\lambda^*(Y, \xi)$ is that this quantity represents the conditional probability of a process $\mathcal{Y}$ having a point in an infinitesimal region containing ξ given the rest of $\mathcal{Y}$ is Y . We can thus compute $d(Y \setminus \xi, \xi)$ for each $\xi \in Y$ and with that we are able to specify all needed components for the BD process. As a part of the BD-MCMC algorithm we will, at each iteration, allow the BD process to run for some finite amount of virtual time T which we can tune so that a sufficient number of BD events occur (typically around 20/iteration). Then, if there are m events in this span such that the last of these happened at $T_m < T$, we pass from the BD process the most recent point configuration $\mathcal{Y}_{T_m}$ to the MCMC portion for model parameter estimation. As well, we can use $|\mathcal{Y}_{T_m}|$ as our estimate for the cardinality of the process in question and thus create a posterior approximation by aggregating these values from the individual iterations.

Appendix B.: Supplementary figures

See Figs. B.6–B.10.

Figure B.6.

Estimated clusterings for various values of the repulsive distance ρ in the Strauss Process for the same simulated point pattern. Using a value of 1, we see overfitting where some clusters are broken apart into several smaller ones, while using 200 leads to a collapsing of multiple clusters into a more regional structure. With the value of 20, we accurately capture the true cluster centers.

Figure B.7.

Histograms of $|C|$, μ_a , η , and Avg. cluster area from the 234 real lung slices analyzed in the exploratory data set (top row) and the 1,800 slices in the analysis set (bottom row).

Figure B.8.

An example simulated point pattern where the BD-MCMC split larger clusters into multiple smaller ones and hence overestimated $|C|$. Black points are estimated cluster centers c_i and black ellipses represent a 75th percentile contour for the covariance matrices Σ_i .

Figure B.9.

Trace plots of $|C|$, η , and Average Area broken down by CT subtype for 200 subjects in analysis data set. Traces go from the bottom of the lung to the top along the x -axis.

Figure B.10.

Example trace plots of $|C|$, μ_α , and η from the BD-MCMC posterior for a simulated data set. The models tend to converge very quickly and mix very well.

Appendix C.: Simulated Data Sets

The general algorithm for generating the simulated point patterns in space S was as follows:

1. Generate sets of cluster centers C from a Strauss process using a spatial BD process

- At each iteration BD process run for a fixed amount of virtual time T so as to have roughly 20 events
 - Process was run for 100 times the number of simulated point patterns needed (e.g. 50,000 to get the 500 simulated data sets for Simulation 1), and then every 100th iteration was saved.
2. Generate observed point patterns for the individual processes $\mathcal{X}_i$
- For each $c_i \in C$, draw $\mu_\alpha \sim N(M_{\mu_\alpha}, V_{\mu_\alpha})$, $\alpha_i \sim N(\mu_\alpha, V_\alpha)$, $\Sigma_i \sim \mathcal{IW}(U, v)$, and $N_i \sim \text{Poisson}(\Lambda_i)$.
 - Draw N_i observed points $x_i \in X_i$ as draws from a $\mathcal{MVN}(c_i, \Sigma_i)$ distribution
3. Generate observed points from random scatter process $\mathcal{X}_0$
- Draw $N_0 \sim \text{Poisson}(\eta \cdot \mathcal{L}(S))$
 - Draw N_0 observed points $x_j \in X_0$ from a uniform distribution on S

Table C.8

Details for the 500 simulated 2D data sets for Simulation 1 and the priors used in the spatial BD-MCMC model fits.

Parameter	Simulated from	Prior in BD-MCMC estimation
C	Inhomogeneous Poisson($B, 17 / \mathcal{L}(B)$) with repulsion parameters $\phi = 1e-9$ and $\rho = 20$	Inhomogeneous Poisson($B, 4 / \mathcal{L}(B)$) with repulsion parameters $\phi = 1e-9$ and $\rho = 20$
$\log(\alpha_i)$	$N(\mu_\alpha, 1)$	$N(\mu_\alpha, 2)$
μ_α	$N(4.7, 1)$	$N(4, 6)$
Σ_i	$\mathcal{IW}(20 \cdot I_2, 20)$	$\mathcal{IW}(40 \cdot I_2, 5)$
η	Fixed at 0.005	Gamma(0.01, 0.01)

Table C.9

Details for 6 parameter combinations used for data set generation in Simulation 2. Here s is used to generate the G matrix for the inverse-Wishart prior on the covariance matrices Σ_i as $W = s \cdot I_2$. The first 3 rows correspond to the real data slices that have the min, median, and max value for $|C|$ (norm), while the second three rows correspond to slices with the min, median, and max values for μ_α . $|C|$ (norm) is computed as $|C| / \mathcal{L}(B)$.

$ C $	$ C $ (norm)	μ_α	η	s
9	0.0003	2.4493	0.0009	8.7455
34	0.0010	3.6920	0.0038	23.2527
34	0.0017	4.9650	0.0005	31.0841
14	0.0005	2.4842	0.0013	11.5653
33	0.0009	3.8210	0.0023	23.8991
29	0.0014	5.1170	0.0004	30.7076

Appendix D.: Visual assessment subtypes

Visual assessment of chest CT scans was conducted In the Quantitative Imaging Lab at National Jewish Health. Subjects were classified into one of ten distinct groups based on the following criteria:

1. **No Airway or Emphysema Abnormality on CT (Normal):**
 - No quantitative emphysema ($\%LAA_{950} < 5\%$)
 - Absent or trace Centrilobular Emphysema (CLE), and absent or trace paraseptal emphysema on visual analysis
 - Bronchial wall thickening absent or mild on visual analysis
 - Absence of significant gas trapping ($\%LAA_{856} < 20\%$)
 - FEV1/FVC ratio > 0.7 , and/or FEV1 $> 80\%$ predicted
2. **Bronchial Airway Disease without Emphysema:**
 - No quantitative emphysema ($\%LAA_{950} < 5\%$)
 - Absent or trace emphysema, and absent or mild paraseptal emphysema on visual analysis
 - Bronchial wall thickening present on visual analysis
3. **Small Airway Disease without Emphysema:**
 - No quantitative emphysema ($\%LAA_{950} < 5\%$)
 - Absent or trace emphysema, and absent or mild paraseptal emphysema on visual analysis
 - Bronchial wall thickening absent or mild on visual analysis
 - One or more indices for airway obstruction
 - Significant gas trapping ($\%LAA_{856} > 20\%$)
 - FEV1/FVC ratio < 0.7 , and/or FEV1 $< 80\%$ predicted
4. **Mild Emphysema:**
 - Absent or mild paraseptal emphysema on visual analysis
 - Mild CLE by Fleischner visual score, or absent or trace CLE by Fleischner visual score with $\%LAA_{950} > 5\%$ and $< 10\%$
5. **Moderate to Severe Emphysema: Upper Lung Predominant:**
 - Absent or mild paraseptal emphysema on visual analysis
 - Upper third/lower third ratio of $\%LAA_{950} > 2$ and either
 - Moderate CLE by Fleischner visual score and $\%LAA_{950} > 5\%$ or

- Confluent or advanced destructive emphysema by Fleischner visual score, and %LAA₉₅₀ > 10%

6. Moderate to Severe Emphysema: Lower Lung Predominant:

- Absent or mild paraseptal emphysema on visual analysis
- Upper third/lower third ratio of %LAA₉₅₀ < 0.5 and either
 - Moderate CLE by Fleischner visual score and %LAA₉₅₀ > 5% or
 - Confluent or advanced destructive emphysema by Fleischner visual score, and %LAA₉₅₀ > 10%

7. Diffuse Moderate to Severe Emphysema:

- Absent or mild paraseptal emphysema on visual analysis
- Upper third/lower third ratio of %LAA₉₅₀ > 0.5 and < 2, and either
 - Moderate CLE by Fleischner visual score and %LAA₉₅₀ > 5% or
 - Confluent or advanced destructive emphysema by Fleischner visual score, and %LAA₉₅₀ > 10%

8. Paraseptal Emphysema (PSE):

- Substantial paraseptal emphysema by visual score (regardless of other emphysema assessments)

9. Discordant – Visual Emphysema without Quantitative Emphysema:

- Absent or mild paraseptal emphysema on visual analysis
- Either of:
 - Moderate CLE by Fleischner visual score and %LAA₉₅₀ < 5% or
 - Confluent or advanced destructive emphysema by Fleischner visual score, and %LAA₉₅₀ < 10%

10. Discordant - Quantitative Emphysema without Visual Emphysema:

- Absent or mild paraseptal emphysema
- Absent or trace emphysema on visual analysis
- %LAA₉₅₀ > 10%

Appendix E.: Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.spasta.2018.12.003>.

References

- Barr RG, Berkowitz EA, Bigazzi F, Bode F, Bon J, Bowler RP, Chiles C, Crapo JD, Criner GJ, Curtis JL, et al., 2012 A combined pulmonary-radiology workshop for visual evaluation of copd: study design, chest ct findings and concordance with quantitative evaluation. *COPD: J. Chronic Obstructive Pulmonary Disease* 9 (2), 151–159.
- Carolan BJ, Hughes G, Morrow J, Hersh CP, O'Neal WK, Rennard S, Pillai SG, Belloni P, Cockayne DA, Comellas AP, et al., 2014 The association of plasma biomarkers with computed tomography-assessed emphysema phenotypes. *Respiratory Res.* 15 (1), 127.
- Castaldi PJ, Cho MH, San José Estépar R, McDonald M-LN, Laird N, Beaty TH, Washko G, Crapo JD, Silverman EK, 2014 Genome-wide association identifies regulatory loci associated with distinct local histogram emphysema patterns. *Am. J. Respiratory Crit. Care Med.* 190 (4), 399–409.
- Cavigli E, Camiciottoli G, Diciotti S, Orlandi I, Spinelli C, Meoni E, Grassi L, Farfalla C, Pistolesi M, Falaschi F, et al., 2009 Whole-lung densitometry versus visual assessment of emphysema. *Eur. Radiol* 19 (7), 1686–1692. [PubMed: 19224221]
- Eddelbuettel D, François R, Allaire J, Chambers J, Bates D, Ushey K, 2011 Rcpp: seamless r and c++ integration. *J. Stat. Softw* 40 (8), 1–18.
- Eddelbuettel D, Sanderson C, 2014 Rcpparmadillo: accelerating r with high-performance c++ linear algebra. *Comput. Statist. Data Anal.* 71, 1054–1063.
- Gu S, Li R, Leader J, Zheng B, Bon J, Gur D, Sciurba F, Jin C, Pu J, 2015 Obesity and extent of emphysema depicted at ct. *Clin. Radiol* 70 (5), e14–e19. [PubMed: 25703460]
- Hame Y, Angelini ED, Hoffman EA, Barr RG, Laine AF, 2014 Adaptive quantification and longitudinal analysis of pulmonary emphysema with a hidden markov measure field model. *IEEE Trans. Med. Imaging* 33 (7), 1527–1540. [PubMed: 24759984]
- Kirby M, Tan WC, Hague CJ, Leipsic JA, Bourbeau J, Hogg JC, Coxson HO, 2017 Computed tomography visual emphysema scoring and quantitative measurements provide independent and complementary information in copd. In: C80-B. MULTI-MODALITY ASSESSMENT OF COPD, ASTHMA, AND ASTHMA-COPD OVERLAP SYNDROME. *Am Thoracic Soc*, pp. A6477–A6477.
- Kong X, Cho MH, Anderson W, Coxson HO, Muller N, Washko G, Hoffman EA, Bakke P, Gulsvik A, Lomas DA, et al., 2011 Genome-wide association study identifies bcl2l1 as a susceptibility gene for emphysema. *Am. J. Respiratory Crit. Care Med.* 183 (1), 43–49.
- Lawson AB, Denison DG, 2002 Spatial Cluster Modelling. CRC press.
- Lynch DA, Al-Qaisi ML, 2013 Quantitative ct in copd. *J. Thoracic Imaging* 28 (5), 284.
- Matin TN, Rahman N, Nickol AH, Chen M, Xu X, Stewart NJ, Doel T, Grau V, Wild JM, Gleeson FV, 2016 Chronic obstructive pulmonary disease: lobar analysis with hyperpolarized 129xe mr imaging. *Radiology* 152299.
- Mendoza CS, Washko GR, Ross JC, Diaz A, Lynch DA, Crapo JD, Silverman EK, Acha B, Serrano C, Estépar R, 2012 Emphysema quantification in a multi-scanner hrct cohort using local intensity distributions. In: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on IEEE, pp. 474–477.
- Mishima M, Hirai T, Itoh H, Nakano Y, Sakai H, Muro S, Nishimura K, Oku Y, Chin K, Ohi M, et al., 1999 Complexity of terminal airspace geometry assessed by lung computed tomography in normal subjects and patients with chronic obstructive pulmonary disease. *Proc. Natl. Acad. Sci.* 96 (16), 8829–8834. [PubMed: 10430855]
- Mohamed Hoessein FA, de Jong PA, Lammers JJ, Mali WPTM, Mets OM, Schmidt M, de Koning HJ, Aalst C.v.d., Oudkerk M, Vliegenthart R, et al., 2014 Contribution of ct quantified emphysema, air trapping and airway wall thickness on pulmonary function in male smokers with and without copd. *COPD: J. Chronic Obstructive Pulmonary Disease* 11 (5), 503–509.
- Møller J, Waagepetersen RP, 2003 Statistical Inference and Simulation for Spatial Point Processes. CRC Press.
- Nakagawa H, Higami Y, Fukunaga K, Uchida Y, Yukimura R, Shigemori W, Kashiwagi Y, Yamaguchi M, Nagao T, Ogawa E, et al., 2016 The effect of emphysema in patients with idiopathic pulmonary fibrosis; evaluation using quantitative ct analysis.

- Nambu A, Zach J, Schroeder J, Jin G, Kim SS, Kim Y-I, Schnell C, Bowler R, Lynch DA, 2016 Quantitative computed tomography measurements to evaluate airway disease in chronic obstructive pulmonary disease: relationship to physiological measurements, clinical index and visual assessment of airway disease. *Eur. J. Radiol* 85 (11), 2144–2151. [PubMed: 27776670]
- Pike D, Mohan S, Ma W, Lewis JF, Parraga G, 2015 Pulmonary imaging abnormalities in an adult case of congenital lobar emphysema. *J. Radiol. case Rep.* 9 (2), 9.
- Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD, 2011 Genetic epidemiology of copd (copdgene) study design. *COPD: J. Chronic Obstructive Pulmonary Disease* 7 (1), 32–43.
- Schroeder JD, McKenzie AS, Zach JA, Wilson CG, Curran-Everett D, Stinson DS, Newell JD Jr, Lynch DA, 2013 Relationships between airflow obstruction and quantitative ct measurements of emphysema, air trapping, and airways in subjects with and without chronic obstructive pulmonary disease. *AJR. Am. J. Roentgenology* 201 (3), W460.
- Sluimer I, Schilham A, Prokop M, Van Ginneken B, 2006 Computer analysis of computed tomography scans of the lung: a survey. *IEEE Trans. Med. Imaging* 25 (4), 385–405. [PubMed: 16608056]
- Stephens M, 2000 Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Ann. Statist* 40–74.
- Vegas-Sanchez-Ferrero G, Washko G, Rahaghi F, Ledesma-Carbayo MJ, Estépar RSJ, 2016 Derivation of a test statistic for emphysema quantification. In: *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on IEEE*, pp. 1269–1273.

Fig. 1.

An example axial slice of lung CT. On the left is a slice masked to only include lung tissue where the gray scale represents the observed HU values. The right panel has LAAs highlighted in red using a -950 HU density mask. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 2.

Progression for an example simulated data set from simulation 1 (top row) and simulation 2 (bottom row). From left to right: 1. Simulated point pattern. 2. Posterior distribution of estimated cluster center locations. 3. Posterior predictive distribution. 4. Estimated cluster locations and size from a single iteration towards the end of the BD-MCMC chain. Black ellipses represent a 75th percentile contour for the covariance matrices Σ_i . Note that in the top row all true clusters have been identified while there is one additional erroneous cluster present in this iteration around the middle-right portion of the slice. This corresponds to the area of the heat map panel that shows a region in this same area that BD-MCMC chain sporadically explored but often immediately deleted the proposed component. In the bottom row we see that the estimated clustering is close to the truth but one small cluster has been missed at this iteration (though is identified on average based on the heat map), while others are collapsed into larger clusters.

Fig. 3.

Progression for two example real CT slices with similar %LAA but varying point process model parameter estimates. From left to right: 1. Simulated point pattern. 2. Posterior distribution of cluster center locations. 3. Posterior predictive distribution. 4. Estimated cluster locations and size from a single iteration towards the end of the BD-MCMC chain. Black ellipses represent a 75th percentile contour for the covariance matrices Σ_i .

Fig. 4. Boxplots of $|C|$, μ_α , η , and Avg. cluster area using subject level means broken down by CT subtype for 200 subjects in analysis data set.

Fig. 5.

Trace plots of μ_α broken down by CT subtype for 200 subjects in analysis data set. Traces go from the bottom of the lung to the top along the x -axis.

Table 1

Components/parameters from the 2 level model and their physical interpretations.

Component	Physical interpretation
$ C $	Number of air holes
C	Locations of air holes
a_i	Size of air hole c_i in terms of number of diseased voxels comprising it
μ_a	Average size of air holes (on log scale)
Σ_i	Measure of dispersion of diseased voxels in air hole c_i from the center

Table 2

Summary of model fits for Simulations 1 and 2.

	RMSE	Rel. RMSE	Bias	Rel. bias	90% credible interval coverage
Simulation 1					
$ C $	1.04	13.94%	0.26	3.5%	87.2%
μ_a	0.31	7.8%	0.17	4.3%	99.8%
η	0.0006	12.62%	0.0001	2.24%	88%
Simulation 2					
$ C $	4.07	19.19%	-3.19	-15.95%	34%
μ_a	0.55	19.2%	0.45	14.15%	93.5%
η	0.0004	49.86%	-0.0001	-16.93 %	81.17%

Table 3

Summary of fitted model parameters from the two sets of CT scans: 234 exploratory CT slices and the 1800 analysis slices. Estimates for $|C|$, μ_a , and average area are normalized by slice area and the values are presented as rates per 20,000 pixels; η is inherently normalized by slice area and is presented as its raw estimate.

	Mean	Std. Dev.	Min	Median	Max
Exploratory CTs					
$ C $	20.62	5.11	6.42	20.22	32.86
μ_a	3.47	0.54	2.22	3.43	5.10
η	0.0027	0.0036	0.0000	0.0013	0.0172
Avg. Area	214.29	94.41	66.91	187.61	573.17
Analysis CTs					
$ C $	18.29	8.95	0	20.38	35.64
μ_a	3.35	0.86	1.41	3.31	8.26
η	0.0024	0.0065	0.0000	0.0003	0.0814
Avg. Area	182.24	77.91	33.35	170.66	669.45

Table 4

Image-based COPD subtypes using visual and quantitative CT assessment.

Subtype code	Description
1	No Airway or Emphysema Abnormality on CT (Normal)
2	Airway Disease without Emphysema (AD): Bronchial Airway Disease (BAD)
3	Airway Disease without Emphysema (AD): Small Airway Disease (SAD)
4	Mild Emphysema
5	Moderate-to-Severe Emphysema: Upper Lobe Predominant
6	Moderate-to-Severe Emphysema: Lower Lobe Predominant
7	Diffuse Moderate-to-Severe Emphysema
8	Paraseptal Emphysema (PSE)
9	Discordant: Visual Emphysema without Quantitative Emphysema
10	Discordant: Quantitative Emphysema without Visual Emphysema

Table 5

Summary of subject level mean estimates for model parameters broken down by CT subtype. The within subject mean is taken over the point estimates (posterior medians) from each of their slices, and the reported means and standard deviations are taken over the subjects in each subtype.

CT Subtype	$ C $		μ_{α}		η		Avg. Area	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
1	8.05	4.85	2.59	0.30	5e-04	8e-04	145.09	52.05
2	8.51	5.11	2.88	0.35	9e-04	0.0014	178.48	46.93
3	10.33	5.19	2.66	0.28	0.0011	0.0013	169.84	62.74
4	17.65	5.83	2.99	0.41	0.002	0.0027	159.58	54.86
5	24.28	3.22	3.82	0.45	0.0012	0.0015	202.65	57.66
6	22.95	3.30	3.84	0.45	8e-04	7e-04	204.07	44.060
7	26.56	3.19	4.14	0.37	0.0013	0.0035	204.78	28.41
8	24.84	5.37	3.85	0.51	0.0021	0.0036	187.66	42.97
9	14.25	3.53	3.05	0.35	5e-04	6e-04	167.04	53.87
10	25.45	2.31	3.64	0.22	0.0133	0.0094	203.25	33.38
Overall	18.29	8.26	3.35	0.65	0.0024	0.0051	182.24	51.87

Table 6

Rank-correlations between four parameters of interest among 200 subjects analysis data set using normalized parameter estimates.

	$ C $	μ_a	η	Avg. Area
$ C $	1.00	0.78	0.12	0.36
μ_a	0.78	1.00	-0.062	0.60
η	0.12	-0.062	1.00	-0.023
Avg. Area	0.36	0.60	-0.023	1.000