Comparing Spatial Point Patterns of Maritime Vessel Locations Using Multivariate Log-Gaussian Cox Processes and Bayesian Inference

Richard Correro
Department of Statistics
Stanford University
rcorrero@stanford.edu

Abstract

In this paper we use a multivariate log-Gaussian Cox Process to model multiple spatial point processes and compare their characteristics. In particular, we use this model to determine whether there is local spatial dependence between two processes. We introduce two datasets, each containing the locations of commercial maritime vessels within two regions (one near the port of Callao and the other containing a region of offshore waters) in the coastal waters of Peru. The first dataset contains the locations of vessels identified in satellite imagery by a computer vision algorithm, and the second contains vessel locations broadcast using Automatic Identification System (AIS) transponders equipped on-board vessels. Data was collected for both regions via both data sources during the month of March in 2020. Using the No-U-Turn Sampler (NUTS), we perform Bayesian inference separately for data corresponding to each target region, in each case yielding posterior density estimates for parameters of interest, one of which determines the degree of local spatial dependence between the two spatial point processes. We analyze the posterior density estimate for this parameter at both target locations and determine that positive local spatial dependence is likely at the location near Callao, whereas a small or negative local spatial dependence is likely in the offshore target region.

1 Introduction

Detecting the location of maritime vessels in littoral, or nearshore, waters is necessary to identify and prevent illegal, unreported, or unregulated (IUU) fishing. Although large vessels are required by international law to be equipped with transponders which broadcast Automatic Identification System (AIS) messages – each of which contains a timestamp, the vessel's latitude and longitude, and its length in meters – these transponders can be disabled by crew [1]. Moreover, small vessels are not required to carry these transponders. For this reason remote, passive methods for accurately identifying the locations of vessels are extremely desirable, and one of the most promising data sources is satellite imagery [2]. Satellite imagery is attractive as it cannot be tampered with unlike AIS and can, in principle, be used to identify the location of even the smallest ships (assuming the imagery is of a sufficiently high resolution). For this reason, deep-learning based computer vision algorithms have been developed to identify vessels, including small vessels, in satellite imagery [2].

The goal of this paper is to compare AIS data and data generated from satellite imagery by a computer vision algorithm, and determine whether ships which appear in one are absent in the other, or vice-versa. For the purposes of this paper, I choose to analyze two regions of ocean within the Peruvian Exclusive economic zone (see figures one and two). The first contains the area immediately surrounding the port of Callao in Lima, Peru, along with the port itself. The second contains a section

STATS 371: Applied Bayesian Inference, Spring 2021, Stanford University, CA.

of littoral waters south of Callao. I also chose to use AIS and satellite imagery data captured during the month of March in 2020.

2 Background

2.1 Point Patterns

A spatial point pattern is a sample of observed spatial locations of things or events [3]. Formally, a point pattern $\mathcal{S}$ is a set of points $\mathbf{s}_i \in \mathbb{R}^2$:

$$\mathcal{S} = \{\mathbf{s}_i : \mathbf{s}_i \in D\}_{i=1}^{N}$$

where $D \subset \mathbf{R}^2$ denotes the region of interest.

For this analysis, we will model the locations of vessels in our region of interest as a spatial point pattern. These locations are generated from two data sources: AIS broadcasts and the outputs of a computer vision algorithm used to analyze satellite imagery of the target area.

2.2 Log-Gaussian Cox Process

A log-Gaussian Cox process (LGCP) is a hierarchical model which combines a Poisson process model at the first level and a Gaussian process at the second [4]. The modelling process begins with a realization $z(\mathbf{s})$ of the Gaussian process, where $\mathbf{s} \in \mathcal{S} \subset D$. Although $z(\cdot)$ has a continuous subspace D as support, we must discretize this domain to make model fitting tractable. We do so by dividing D into a set of M grid cells and use each cell's centroid as a representative point for $z(\cdot)$ within that region. That is, we approximate $z(\cdot)$ over D by estimating

$$\{z(\mathbf{c}_j)\}_{j=1}^{M}$$

where $\mathbf{c}_j$ denotes the location of the centroid of grid cell A_j. We also define w_j as the area of A_j. Increasing M leads to better approximations of $z(\cdot)$ but comes at the expense of increased computational cost. For the purposes of this analysis, we assume

$$z(\mathbf{s}) \sim GP(\mathbf{0}, \mathbf{C})$$

where $\mathbf{C} \in \mathbb{R}^{M \times M}$ denotes a valid covariance matrix between centroids [5]. We denote by $\lambda(\mathbf{c_j})$ the intensity associated with A_j, where

$$\lambda(\mathbf{c_j}) = \exp\{z(\mathbf{c}_j)\}.$$

With this, we treat X_j, the number of points within A_j, as a Poisson random variable:

$$\begin{aligned} X_j &\sim \text{Poisson}\left(\int_{A_j} \lambda(\mathbf{s}) d\mathbf{s}\right) \\ &\approx \text{Poisson}(w_j \cdot \lambda(\mathbf{c}_j)). \end{aligned}$$

2.3 Multivariate Extension of the Log-Gaussian Cox Process

Although the log-Gaussian Cox Process is extremely useful, we need to model two separate datasets as point processes over D, which requires extending the LGCP model. We do so by treating $\boldsymbol{\lambda}(\mathbf{s})$ as a vector of field intensities

$$\boldsymbol{\lambda}(\mathbf{s}) = (\lambda_1(\mathbf{s}), \ldots, \lambda_P(\mathbf{s}))$$

where P denotes the number of spatial point processes being modelled. In this way we may model the spatial intensity for point pattern $\mathcal{S}_p$ by the pth component of $\boldsymbol{\lambda}(\mathbf{s})$. We model the $z_1(\mathbf{s}), \ldots, z_P(\mathbf{s})$ by

$$\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), \ldots, z_P(\mathbf{s}))^T$$

which is a multivariate Gaussian process with mean $\mu = (\mu_1, \ldots, \mu_P)^T$ and covariance functions

$$k_{p,p'}(\mathbf{s}, \mathbf{s}') = Cov(z_p(\mathbf{s}), z_{p'}(\mathbf{s}')), \qquad p = 1, \ldots, P,\ p' = 1, \ldots, P.$$

Under this model, the point patterns $\mathcal{S}_1, \ldots, \mathcal{S}_P$ are treated as the results of Poisson processes which are conditionally independent given $\boldsymbol{\lambda}(\cdot)$. Each process p is associated with an intensity function $\lambda_p(\mathbf{s})$ which is defined over D.

3 Model

3.1 MLGCP in Detail

The multivariate log-Gaussian Cox process (MLGCP) may be converted to a set of linear equations involving only univariate Gaussian processes (for more about this approach, see [5], specifically pp.103-110). By modeling $z_p(\cdot)$ as a linear function of these processes, we may capture any spatial dependencies which exist between spatial point processes in an easy-to-interpret manner [5].

Let

$$G_p(\mathbf{s}), \qquad p = 1, \ldots, P$$

denote a set of univariate Gaussian processes each with mean μ_p and variance 1. Further let $c_p(\mathbf{s}, \mathbf{s}')$ denote a valid covariance function corresponding to process p. Then we have

$$z_p(\mathbf{s}) = \sum_{p'=1}^{P'} \alpha_{p,p'} G_{p'}(\mathbf{s}),$$

where $\alpha_{p,p'} \in \mathbb{R}$. By sharing this set of Gaussian processes between the intensity functions associated with each spatial point process, this model may capture any spatial dependencies which exist between the point processes [5]. In practice we will treat $\alpha_{p,p'}$ as an unknown parameter and estimate its value using Bayesian inference. This coefficient determines the degree of local spatial dependence between our spatial point processes [5]. Here "local dependence" is referring to dependence across the two datasets at a given location within the target region D. If, for some $\mathcal{S}_p$, $\mathcal{S}_{p'}$ where $p \neq p'$, our posterior estimate for $\alpha_{p,p'}$ places sufficient probability on positive values, then we have good reason to believe that a positive local dependence exists between the two samples. If, on the other hand, the posterior credible interval includes 0, then it is likely that there is no local spatial dependence between the samples. And finally, if the posterior places sufficient probability on negative values, this would suggest that there is a negative dependence between the samples, i.e. if there are points from one sample in a grid cell then it is less likely that there will be points from the other sample there, and vice-versa.

3.2 Fitting the MLGCP

To fit a multivariate log-Gaussian Cox process using Bayesian inference we must define a set of prior distributions corresponding to the unknown parameters along with a likelihood function. We will discuss the prior densities later.

For the likelihood function, given spatial point patterns $\mathcal{S}_1, \ldots, \mathcal{S}_P$, we have

$$f(\boldsymbol{\lambda}(\cdot)|\mathcal{S}_1, \ldots, \mathcal{S}_P) = \prod_{p=1}^{P} \prod_{\mathbf{s}_j \in \mathcal{S}_p} \lambda_p(\mathbf{s}_j) \exp\left(\int_D \lambda_p(\mathbf{s}) d\mathbf{s}\right)$$

[5]. Again, we must approximate this integral by dividing D into a set of M grid cells.

The primary reason for using this model is that it allows for estimation of the local spatial dependence between the two samples. However, this model also allows for analysis of the spatial dynamics of each population individually by way of the intensity functions [5].

3.3 Details About the Data

We use the multivariate log-Gaussian Cox process to analyze the spatial dynamics of fishing and commercial vessels in the coastal waters of Peru. For the purposes of this analysis, two target locations were chosen as was a start date and end date for data collection. We will refer to the first location, which contains a region of waters immediately surrounding the port of Callao in Peru, along with the port itself, as the "Callao" region. The second region, which covers an area of the open ocean between approximately 50 and 15 kilometers off the coast of Peru, will be referred to as the "offshore" region. I collected two samples of spatial point process data corresponding to the locations of commercial vessels in both regions (see figures 3-6). The AIS data was obtained from Global Fishing Watch, a nonprofit organization which captures AIS data from multiple locations

globally. The satellite imagery data was generated using a bespoke computer vision algorithm which I am developing as part of my research for the Center for Ocean Solutions in the Stanford Woods Institute (see here for more details about this model). I chose to include only data corresponding to the month of March 2020 and chose these particular target locations because they were likely to exhibit dissimilar spatial dynamics.

The Callao region covered the waters immediately surrounding the port of Callao in Lima, Peru. This is a very high density region for vessels and includes several anchorages, places where vessels may linger for days at a time. Importantly, this port is frequented predominately by larger commercial ships which are required by international law to utilize AIS transponders [1]. Because of this, we would expect that many of the vessels identified in the satellite imagery would appear in the AIS dataset as well.

The offshore region is south of Callao and abounds with anchoveta during the month of March. Because of this many fishing vessels, both large and small, will visit this location every March to engage in fishing. Although some of these vessels may be equipped with AIS transponders, it is likely that many of these vessels are sufficiently small to be exempt from international and Peruvian regulation which mandates the use of AIS [2]. Further, this region includes waters in which commercial fishing vessels are not allowed to operate, and it is possible that some larger fishing vessels which wish to obscure their illegal fishing will unlawfully disengage their AIS transponders [2]. For these two reasons we have reason to believe that the distribution of vessels observed in satellite imagery may be significantly different for the distribution of vessels observed in the AIS data in this region.

For the purposes of this analysis, we are only interested in identifying whether the distributions of each dataset demonstrate local spatial dependence. As spatial dependence is a symmetric property of the two spatial point processes, we can only model the relative dependence between the datasets. However, the AIS data is broadcast much more frequently (on the order of seconds) than satellite imagery is acquired (generally on the order of days for any one location), and therefore the absence of ships in satellite imagery which are present in AIS is to be expected. The reverse, however, is of interest as this has implications for determining whether IUU activity is occurring within a region.

3.4 Formal Definition of The Problem

Given a target region D, let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the spatial point processes corresponding to our two datasets within D. $\mathcal{S}_1$ denotes the locations of vessels as determined through the use of a computer vision algorithm which automatically identifies ships within satellite imagery, and $\mathcal{S}_2$ denotes the locations of vessels in data received from AIS transponders. We model the locations of vessels in $\mathcal{S}_1$ and $\mathcal{S}_2$ using a multivariate log-Gaussian Cox process with joint intensity

$$\boldsymbol{\lambda}(\mathbf{s}) = (\lambda_1(\mathbf{s}), \lambda_2(\mathbf{s}))^T$$

where $\lambda_1(\mathbf{s})$ is the intensity function corresponding to the satellite imagery data and $\lambda_2(\mathbf{s})$ corresponds to the AIS data.

We define

$$\log \boldsymbol{\lambda}(\mathbf{s}) = \mathbf{z}(\mathbf{s})$$

where $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), z_2(\mathbf{s}))$ satisfies

$$\begin{aligned} z_1(\mathbf{s}) &= \alpha_{1,1} G_1(\mathbf{s}) \\ z_2(\mathbf{s}) &= \alpha_{2,1} G_1(\mathbf{s}) + \alpha_{2,2} G_2(\mathbf{s}). \end{aligned}$$

for univariate Gaussian processes $G_1(\mathbf{s}),\ G_2(\mathbf{s})$ with means $\mu_1,\ \mu_2$, respectively, and variances equal to one. Here $G_1(\mathbf{s}),\ G_2(\mathbf{s})$ are realizations from independent Gaussian processes at the M centroids corresponding to the M grid cells of D.

For each Gaussian process we use an exponential covariance function of the form

$$c_p(\mathbf{s}, \mathbf{s}') = \exp\left\{\frac{||\mathbf{s} - \mathbf{s}'||}{\phi_p}\right\}$$

for $p = 1,\ 2$. The parameter ϕ_p controls the rate of decay of spatial correlation as a function of distance.

Table 1: Prior Distributions for Model Parameters

$\alpha_{2,1}$	$\mathcal{N}(0, 100)$
$\alpha_{1,1}$	InvGamma(shape $= 2$, scale $= 2$)
$\alpha_{2,2}$	InvGamma(shape $= 2$, scale $= 2$)
μ_1	$\mathcal{N}(0, 1)$
μ_2	$\mathcal{N}(0, 1)$
ϕ_1	Unif$(0.005, 0.1)$
ϕ_2	Unif$(0.005, 0.1)$

3.5 Priors

This model contains seven unknown parameters which must be estimated using Bayesian inference. These parameters, along with their corresponding prior distributions, are listed in table 1. Note: $\mu_1,\ \mu_2$ denote the means of Gaussian processes $G_1,\ G_2$, respectively.

4 Algorithm

We perform model inference using a Hamiltonian Monte Carlo algorithm. Specifically, we use the No-U-Turn Sampler (NUTS) for posterior sampling [6]. This algorithm was chosen for its speed and accuracy. We implemented our model in Python using the PyMC3 package [7].

5 Results

I performed two separate tests, fitting the model separately for each of the two target region and using the corresponding datasets for each location. I divided each target region into $M = 100$ grid cells and summed the number of vessels contained in each cell for the AIS data. I then repeated the process with the satellite imagery data.

For each target location, the posterior was sampled 2000 times of which the first 1000 were discarded and the last 1000 retained for analysis. For each location posterior means and 94% highest density intervals are provided for the seven parameters (see figures 9 and 10). Trace plots for each parameter are also provided (see figures 7 and 8).

For the purposes of this paper, the most interesting parameter is $\alpha_{2,1}$ as its posterior distribution can reveal whether there is spatial dependence between the datasets [5]. In the data corresponding to the Callao region, we see that the 94% highest density region (HDI) of the posterior density for $\alpha_{2,1}$ lies between 1.8 and 13 (see the panel labeled "alpha_21" in figure 9). Because of this, we may conclude that there is a positive local spatial dependence between satellite imagery data and AIS data in the Callao region.

For the data corresponding to offshore region, however, the 94% HDI ranges from -10 to 4.2 (see the panel labeled "alpha_21" in figure 10). Based on this, we may conclude that the spatial dependence between the datasets in this region is either very small or negative.

6 Conclusion

In this paper I began by describing the univariate log-Gaussian Cox process, a model for spatial point pattern data which treats the locations of data within a target region as a random variable. I then introduced a multivariate extension of this model which we used to model multiple spatial point processes and compare their characteristics. In particular, this model may be used to determine whether there is local spatial dependence between two processes. I introduced two datasets, each containing the locations of commercial maritime vessels within two regions in the coastal waters of Peru. These regions were chosen based on the assumed spatial dynamics exhibited by vessels within those regions. The first dataset contains the locations of vessels identified in satellite imagery by a computer vision algorithm, and the second contains vessel locations broadcast using Automatic Identification System (AIS) transponders equipped on some vessels. Data was collected for both regions via both data sources during the month of March in 2020. Using NUTS, a Hamiltonian

Monte Carlo sampling algorithm, I performed Bayesian inference separately for data corresponding to each target region, in each case yielding posterior density estimates for seven parameters of interest. One of these, $\alpha_{2,1}$ determines the degree of local spatial dependence between the two spatial point processes, one corresponding to vessel locations contained in the satellite imagery and the other to vessel locations identified using AIS within the target region. We analyzed the posterior for $\alpha_{2,1}$ at both target locations and determined that positive local spatial dependence is likely at the location near Callao, whereas a small or negative local spatial dependence is likely in the offshore target region.

References

[1] Fréon, Pierre, et al. "Environmentally extended comparison table of large-versus small-and medium-scale fisheries: the case of the Peruvian anchoveta fleet." Canadian Journal of Fisheries and Aquatic Sciences 71.10 (2014): 1459-1474.

[2] Kanjir, Urška, Harm Greidanus, and Krištof Oštir. "Vessel detection and classification from spaceborne optical images: A literature survey." Remote sensing of environment 207 (2018): 1-26.

[3] Baddeley, Adrian, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. CRC Press, (2016).

[4] Ming Teng, Farouk Nathoo and Timothy D. Johnson. "Bayesian computation for Log-Gaussian Cox processes: a comparative analysis of methods, Journal of Statistical Computation and Simulation", 87:11, 2227-2252, (2017).

[5] Gelfand, Alan E., and Erin M. Schliep. "BAYESIAN INFERENCE AND COMPUTING FOR SPATIAL POINT PATTERNS." NSF-CBMS Regional Conference Series in Probability and Statistics 10 (2018).

[6] Hoffman, Matthew D., Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." Journal of Machine Learning Research, 15:47, 1593-1623, (2014).

[7] Salvatier J., Wiecki T.V., Fonnesbeck C. "Probabilistic programming in Python using PyMC3". PeerJ Computer Science 2:e55 DOI: 10.7717/peerj-cs.55. (2016).

Project code available here, and vessel detection algorithm code available here.

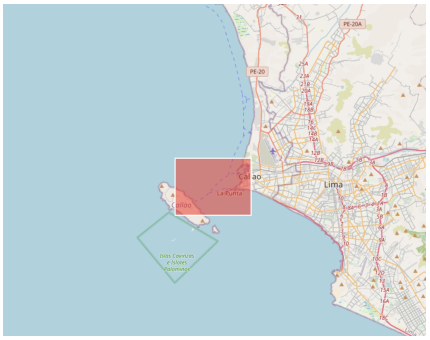

Figure 1: Callao target region.

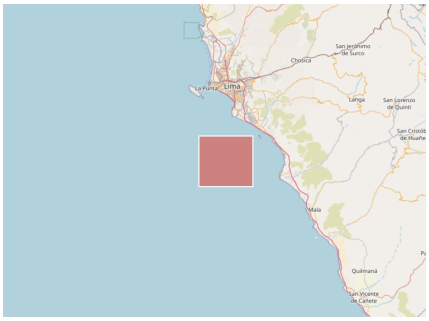

Figure 2: Offshore target region.

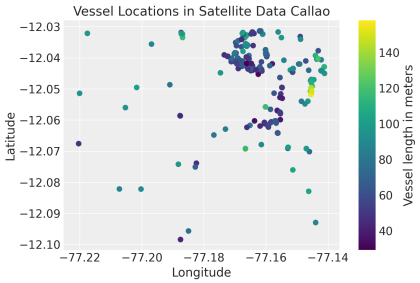

Figure 3: Vessel locations in Callao satellite data.

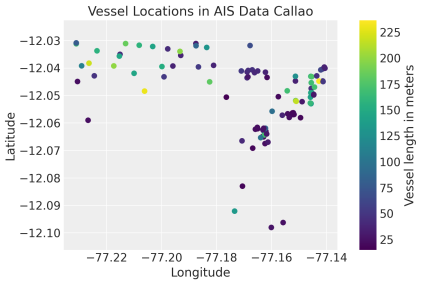

Figure 4: Vessel locations in Callao AIS data.

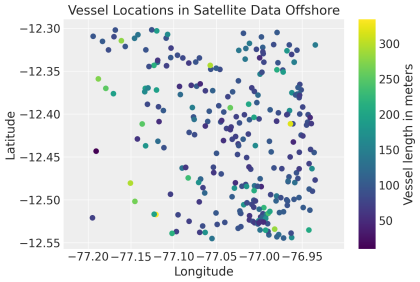

Figure 5: Vessel locations in offshore satellite data.

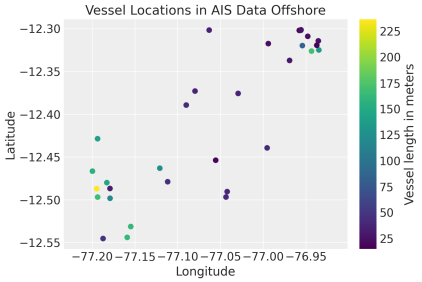

Figure 6: Vessel locations in offshore AIS data.

Figure 7: Traceplots for Callao model.

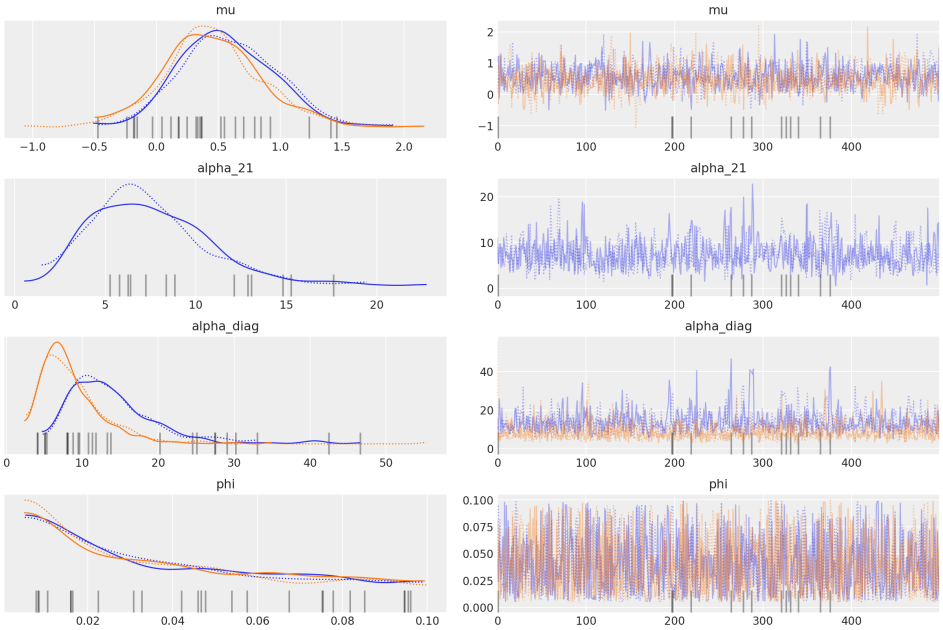

Figure 8: Traceplots for offshore model.

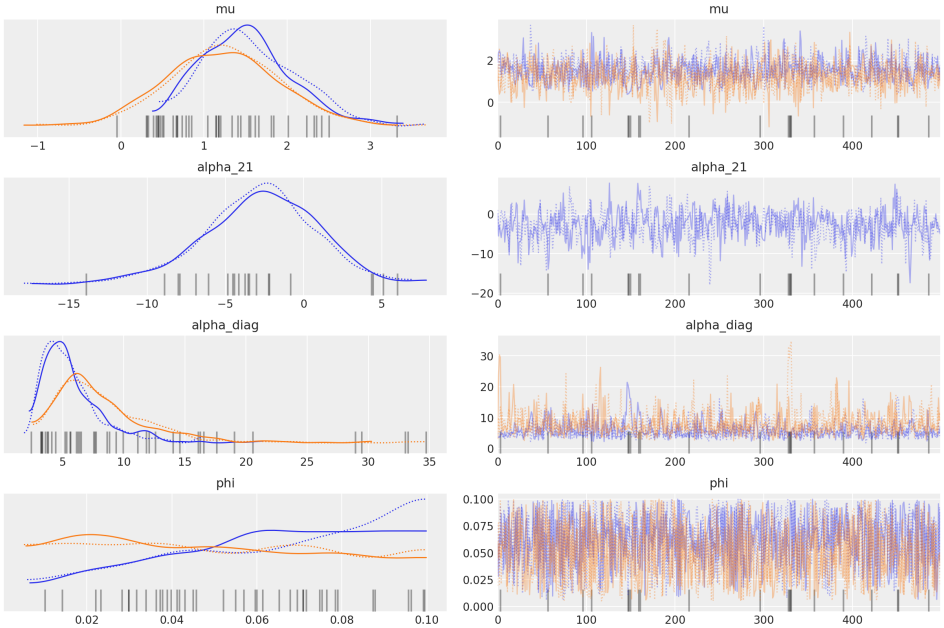

Figure 9: Posterior density estimates for Callao model.

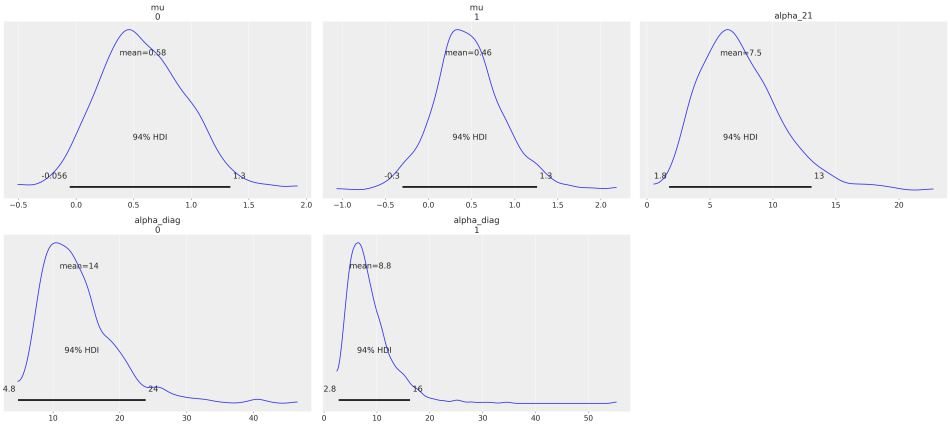

Figure 10: Posterior density estimates for offshore model.