# Estimating the average number of microplastic particles per liter in the world's oceans using spatial statistics techniques

By

Kathryn Konrad

Senior Honors Thesis

Department of Statistics and Operations Research

University of North Carolina at Chapel Hill

15 April 2021

Approved:

_____

Richard L. Smith, Thesis Advisor

Vladas Pipiras, Reader

# Contents

## Abstract

As attention is paid to the state of plastic pollution in the environment, an increasing area of research is in ocean microplastics. These very small pieces of plastic can come in a variety of shapes and colors and occur when larger plastic pieces begin to decay. Microplastics have been found on land, in surface water, in the depths of the ocean, and even in the bellies of fish. They are an important subset of plastic pollution, and one that could easily work through the ocean ecosystem with unknown consequences. While there is not a large amount of data on the subject, this paper attempts to model the spatial distribution of these microplastics in the world's oceans. The data used was collected by the group Adventure Scientists, an organization that relies on citizen scientists to collect data on a variety of environmental issues. Collected from 2013 to 2017, there were 1393 points across the oceans. The spatial statistics methods used to calculate the estimate of the average number of particles per liter was kriging. The average amount of microplastics per liter in the ocean is estimated to be 5.4757 pieces per liter. Pieces are highly concentrated in the polar regions and other areas of accumulation. This paper expands on research done by staff at Adventure Scientists by including spatial statistics methods.


## I. Overview/Purpose

Plastic is an incredible material. It is durable, moldable, flexible, lightweight, and can be used for an incredibly vast array of products, from playset slides to single-use water bottles to casing for medical monitoring tools. Unfortunately, their adaptability also means that plastics can be found almost ubiquitously, including in the oceans. While some plastics are recycled and more still make their way into landfills, plastic waste that is not responsibly disposed of can be seen in staggering quantities in the oceans. According to the nonprofit group the Ocean Conservancy, there are an estimated 150 million tons of plastic in the ocean, with 8 million tons entering every year (Leonard, 2020). This plastic makes its way throughout the ocean ecosystem, being mistaken as food by animals such as sea turtles, fish, and seagulls, as well as absorbing harmful chemicals.

While it seems like plastic is indestructible, it does break down into what are called microplastics. Microplastics are generally considered to be pieces of plastic ranging up to about 5 millimeters, though this definition is flexible. Microplastics occur when wear and tear break down a plastic, or when small fibers are shed from synthetic or semisynthetic materials. These pieces are called microfibers and will be referred to as "filament-shaped" through the rest of this report. Shockingly, a 2018 study on the frequency of microplastics in mesopelagic fish in the northwest Atlantic ocean found 73% of the sampled fish to have microplastics and microfibers in their bellies, a statistics that should concern

anyone (Wieczorek et al, 2018). Researchers have even found microplastics in sea salt produced for human consumption, between 50-280 microplastic pieces per kilogram of salt (Iñiguez, 2017).

If any progress is to be made to clean up marine plastics and microplastics, the spatial distribution would be incredibly useful to know. If it is known where microplastics tend to form and collect, then groups that clean up ocean waste could concentrate their efforts on those critical areas, thus optimizing their results. Plastics may collect in the polar regions or in gyres. Gyres are regions in the middle of an ocean formed by swirling currents. These currents are usually along the equator and along coastlines, and they sweep floating plastics or plants into the center of the gyre. The floating pieces then accumulate in the center of the gyre. The Sargasso Sea off the coast of North America, a region of slow-moving waters filled with the algae sargassum, is the result of a gyre, as is the Great Pacific Garbage Patch off the coast of California, which is a massive accumulation of marine debris. There are five gyres in the world's oceans: the Northern Pacific, the Southern Pacific, the Northern Atlantic, the Southern Atlantic, and the Central Indian gyres. The remainder of this paper focuses on the data collection effort and the statistical components for modeling the marine microplastic spatial distribution.

## II. Data Gathering

Data collection efforts began in May of 2020. Internet search engines were used first in an effort to find a dataset of ocean plastics on a global scale, as well as several experts in the field. These sources did provide some relevant datasets, which are listed below in the table. Many of the datasets were collected by citizen science groups. Datasets tended to include the date of sampling, location (longitude and latitude), data about the sea surface conditions, including Beaufort scale rating and surface salinity at the time of sampling, and some measurement of the number of microplastics in the sample. This final measure varied the most, and this lack of standardization is part of the difficulty with finding consistent microplastics data. Also, there were a fair number of studies that collected information of plastic waste but did not include microplastics.

| *Organization/Individual Owner* | *Years collected* | *Number of datapoints* | *Measurement of microplastics* | *Reason for not using* |
|---|---|---|---|---|
| NOAA boat race | 2017-2018 | 96 | Particles per cubic meter | 96 data points was not a sufficient amount to complete a global spatial model. |

| | | | | |
|---|---|---|---|---|
| Mississippi State University | 2017-2018 | 590 | Number of particles | While this dataset contained a fair amount of data, the geographic distribution was not wide enough for this project. It covered only the coastal areas of the Gulf of Mexico. |
| L. Lebreton and M. Eriksen | 2007-2013 | 1571 (680 capturing micro-plastics) | Particles per square kilometer | This dataset has more data points than the dataset used. However, it was collected using surface manta tows, which have been known to undersample microplastics (Barrows, 2017). |
| Adventure Scientists | 2013-2017 | 1393 | Particles per liter | This dataset included a fair amount of data, including a breakdown of the pieces by shape and color. Its geographic distribution was one of the broadest. |

## Dataset of Interest

After exploring the available data about microplastics, the dataset collected by the group Adventure Scientists was chosen for the spatial modeling. Adventure Scientists is an organization dedicated to collecting data about environmental issues. They tend to use volunteers and outdoor enthusiasts to collect data from remote areas, provide training for these volunteers, assemble the data, and provide quality control checks for their data before making it available. This particular dataset was collected from 2013 to 2017 by volunteers. Volunteers would either be already going on a boat charter or would join a charter already planned. Thus, the samples are not random and were not intentionally designed by Adventure Scientists. It includes 1393 points across all five oceans, making it one of the most extensive datasets specifically about microplastics to date. The volunteers would collaborate with boats that were already going out to sea, and samples were taken throughout the voyage. Buckets of

approximately one liter in volume were used to collect seawater, and the location, data and time, sea surface salinity, wind speed, and other variables were recorded for each sample. The microplastics themselves were counted and classified back on land. The process was done with a high level of quality assurance in mind and can be read about in detail in the 2018 paper by Barrows et al.

Some of the drawbacks of this dataset were the lack of a sampling scheme and the varying density of samples across the globe. Because samples were taken from pre-existing boat charters, they are not randomized and do not conform to a grid sampling method, which would be preferable for spatial modeling. It would be preferable if the dataset included samples taken from randomly-selected geographic coordinates. However, this is logistically unfeasible given the remoteness of the open ocean. Further, some areas such as the mid-Atlantic Ocean have a much higher sampling density of points while other areas like the Southern Ocean and the Indian Ocean have much fewer points. This is not ideal for the aim of this project, which is to study the global distribution of microplastics in the ocean. If one part of the global ocean is undersampled, then this will weaken the "globalness" of the results. Neither of these traits are ideal for the spatial modeling, but this dataset was one of the few with over a thousand samples and specifically focused on microplastics.

One region well-sampled in the dataset is the mid-Atlantic. This is because Adventure Scientists partnered with the Atlantic Rally for Cruisers race, a trans-Atlantic boat race that took place between November and December 2014. 473 samples were collected through this collaborative effort, accounting for a large percentage of the open ocean data.

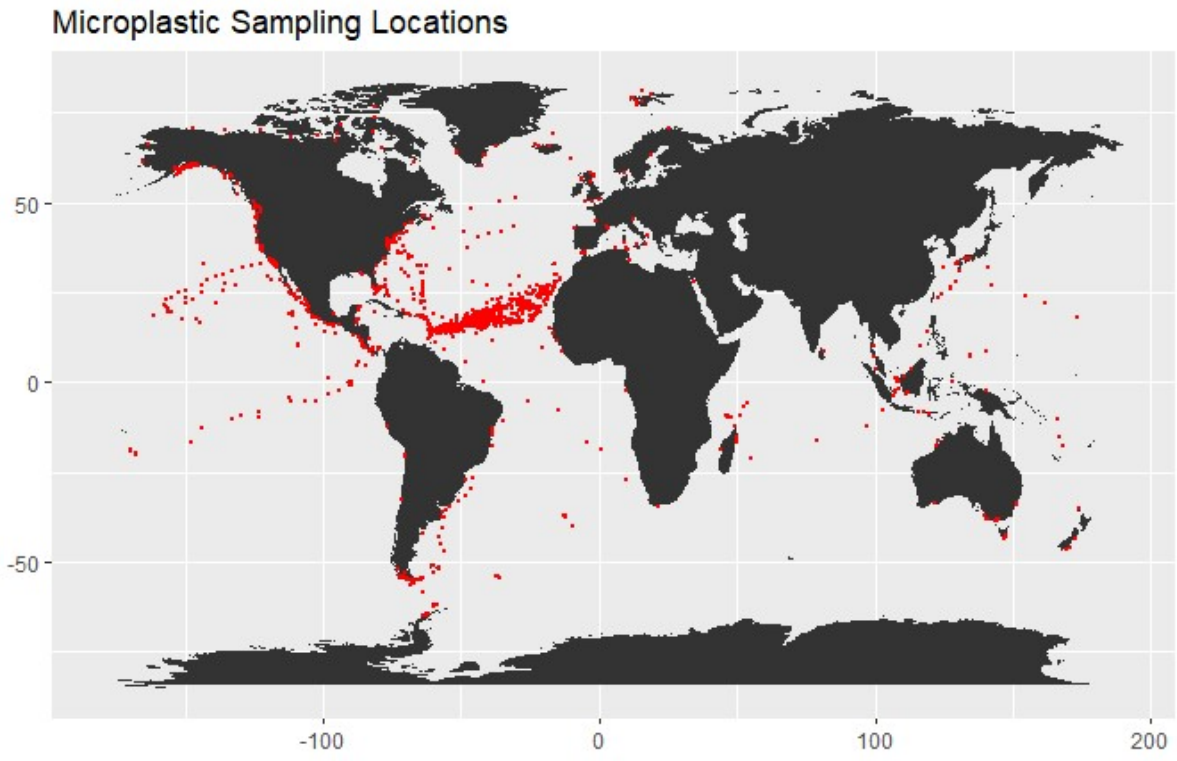
The other dataset with over a thousand samples was the Lebreton dataset. Collected from 2007-2013 and with 1571 datapoints, there are more samples than the Adventure Scientists dataset. This dataset has similar drawbacks as the Adventure Scientists dataset: the sample locations are not random, and the distribution of samples is varying in density across the globe. In addition to these, the Lebreton dataset included two sampling methods: a surface trawl and a visual survey. The surface trawl method involved dragging a net across the ocean surface for a length of time between 15 and 60 minutes. The visual survey was intended to record occurrences of large plastics such as discarded fishing net or plastic bottles. Observers recorded the number and type of large plastics they saw off the bow of the ship for a length of time between 15 and 60 minutes. There were 680 surface trawl datapoints and 891 visual survey datapoints. This dataset was not used because only the 680 surface trawl datapoints can report on the microplastic distribution.

This dataset includes 1393 data points and 141 variables. The majority of those variables are subsets of the total microplastics counts, broken down by color, size, and type of plastic. The possible colors were blue, black, red, green, transparent, and “other”, the possible shapes were round-shaped, filament-shaped, and “other”, and the possible sizes were less than 1.5 millimeters, 1.6-3.1 millimeters, 3.2-5 millimeters, and 5.1-9.6 millimeters. Additional covariates collected were the date and time of the sample, the latitude and longitude of the sampling location, as well as several written location descriptors (such as which ocean or whether the location was coastal or in the open ocean), the water temperature in degrees Celsius, depth of the sample in meters, the wind speed in knots and wind direction, the date filtered, the date counted, and sample volume.

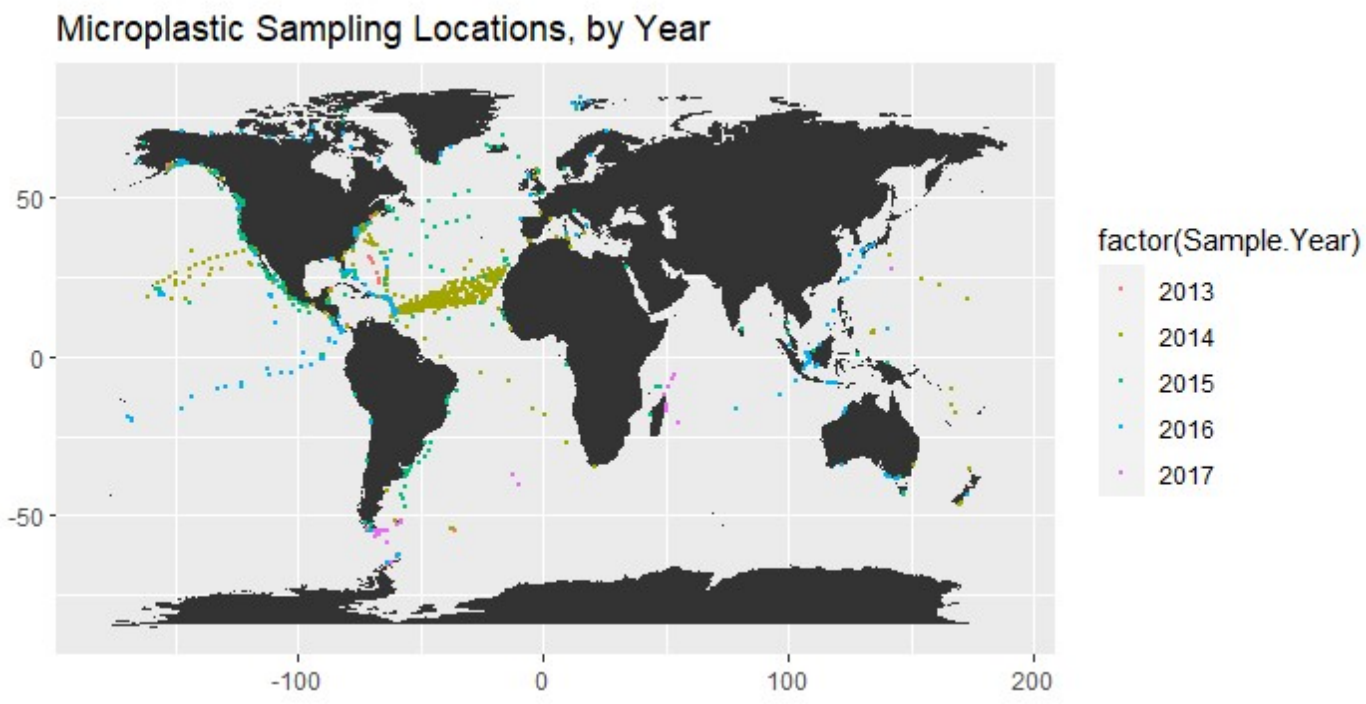
Previous work with this dataset was done by researchers from Adventure Scientists. In the 2018 paper about the dataset, Barrows describes the overall patterns seen in the data. Microfibers are the most commonly seen particles, and the highest concentrations of particles are found in the polar regions, as anticipated by several models (Isobe, 2017; Wilcox, 2015). Generally, the open ocean had higher concentrations than coastal regions. The global microplastic average based off the data was found to be 11.8 particles per liter with a standard deviation of 24 particles per liter, an estimate roughly three times higher than other studies (Barrows, 2018). Barrows believes this is because the grab sampling method allows for more pieces to be caught and counted.

## III. Exploratory Data Analysis

Data was collected from across the globe, but particularly in the mid-Atlantic Ocean and along the coastline of North America. The map below shows the locations of the data samples, with each red point indicating one sample.



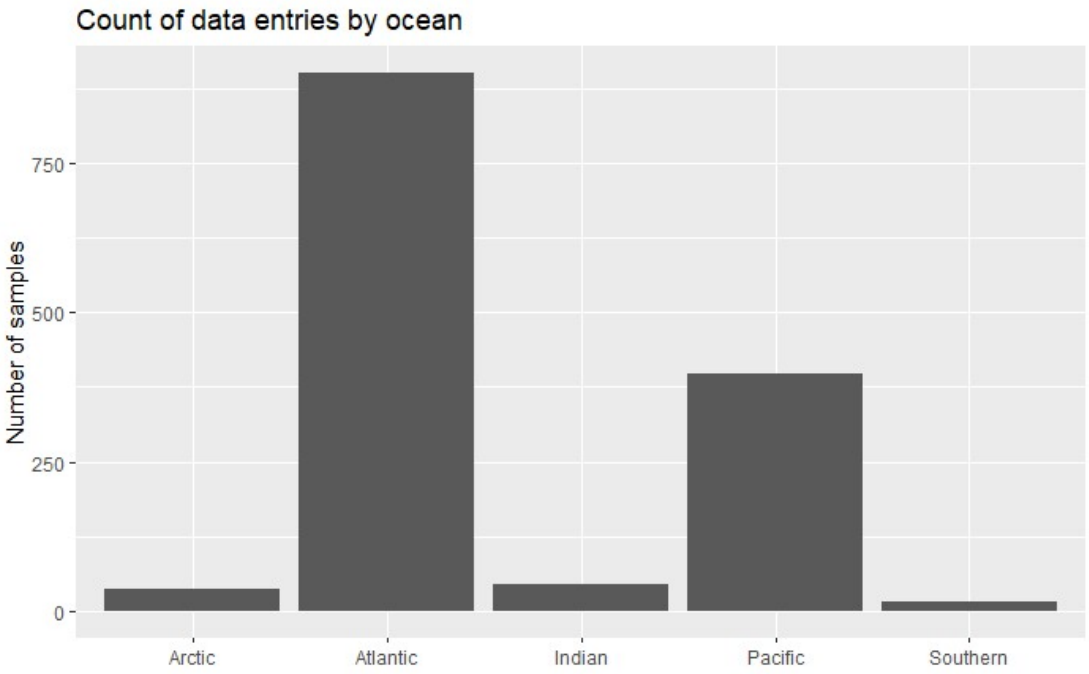

When looking at the data by location and year, we can see that many of the visually grouped clusters were sampled during the same year. For example, the large Atlantic swath was all sampled during 2014 while the points off the coast of Japan were sampled during 2016. This should hopefully make analysis within each data-area possible without too much worry about the temporal influence. Of course, the data could be further broken down and analyzed while considering the temporal ocean changes, and we will look for evidence of spatial correlation within each year later in this document. The number of samples in each year is noted below the map, as well as where they tend to be located.



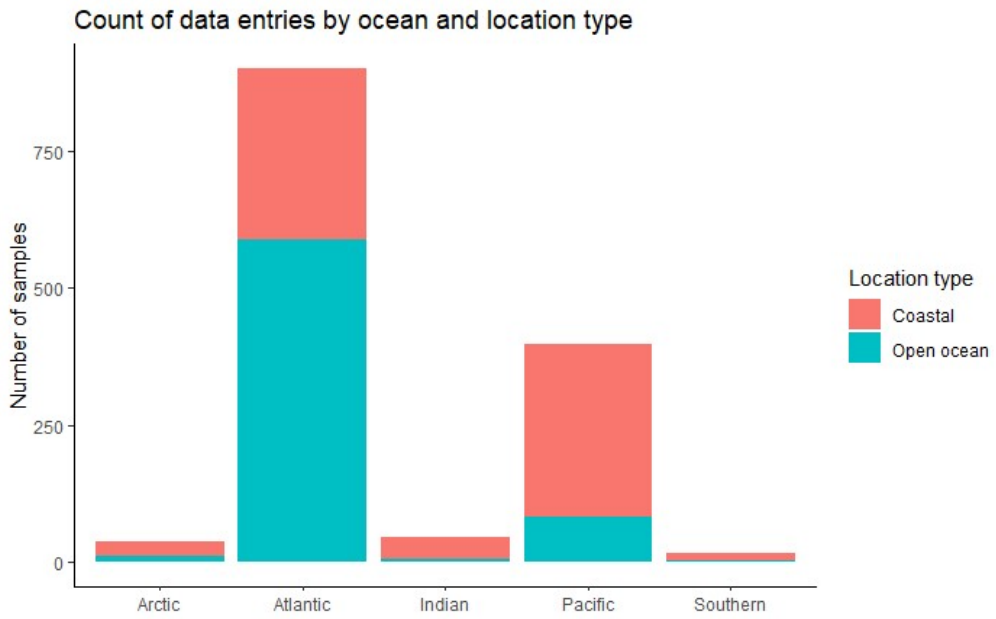

| Year | Number of points | Regions |
|---|---|---|
| 2013 | 50 | West central Atlantic, Southern Alaska, Southern Atlantic/Antarctica |
| 2014 | 725 | Central Atlantic, Great Pacific Garbage Patch, East and West coasts of North America, Mediterranean |
| 2015 | 318 | Central Atlantic, West coast of North America, East coast of South America |
| 2016 | 249 | North of North America, East Pacific, Southeast Asia, Caribbean |
| 2017 | 51 | Southern tip of South America/Antarctica, Madagascar/Indian Ocean |

As we can see from both the map of location points and the histogram of Ocean Basins (counts of records in each ocean), there are the most points in the Atlantic Ocean, followed by the Pacific, the Indian, the Arctic, and the Southern. We can see from the map that there is a large swath of points in middle of the Atlantic Ocean (20 degrees North), looking to be the densest part of the map. The coasts of North America also seem to be covered fairly well. There is a noticeable gap of data in the Southern

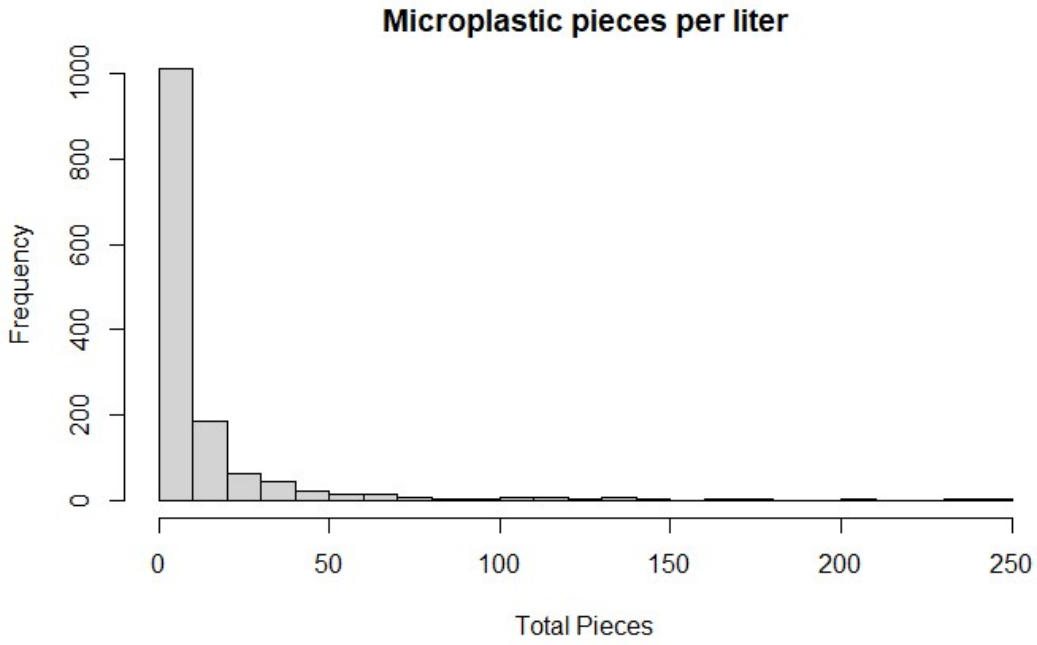
Ocean, the South of the Indian Ocean, and the South of the Pacific Ocean. This is not ideal when attempting to model the microplastic distribution globally but must be tolerated.




To go further into where samples were taken, we can look at whether the samples were taken from a coastal location or an open ocean location. Each data points of the dataset includes this variable, which reflects how far the point is from the nearest shoreline. If the point is within 12 nautical miles from shore, it is classified as coastal; otherwise, it is classified as open ocean. This distance is what the United Nations Convention on the Law of the Sea defines as the breadth of a state's territorial sea (UNCLOS, 1982) and was used by Adventure Scientists as the demarcation between the two states (Barrows, 2018). Generally, there is more coastal data than open ocean data for each ocean, though the Atlantic Ocean has more open ocean data. There are 708 coastal locations and 685 open ocean locations represented in the data. This variable could be included as a covariate in the model, but as of the writing of this report, it has not yet been included.

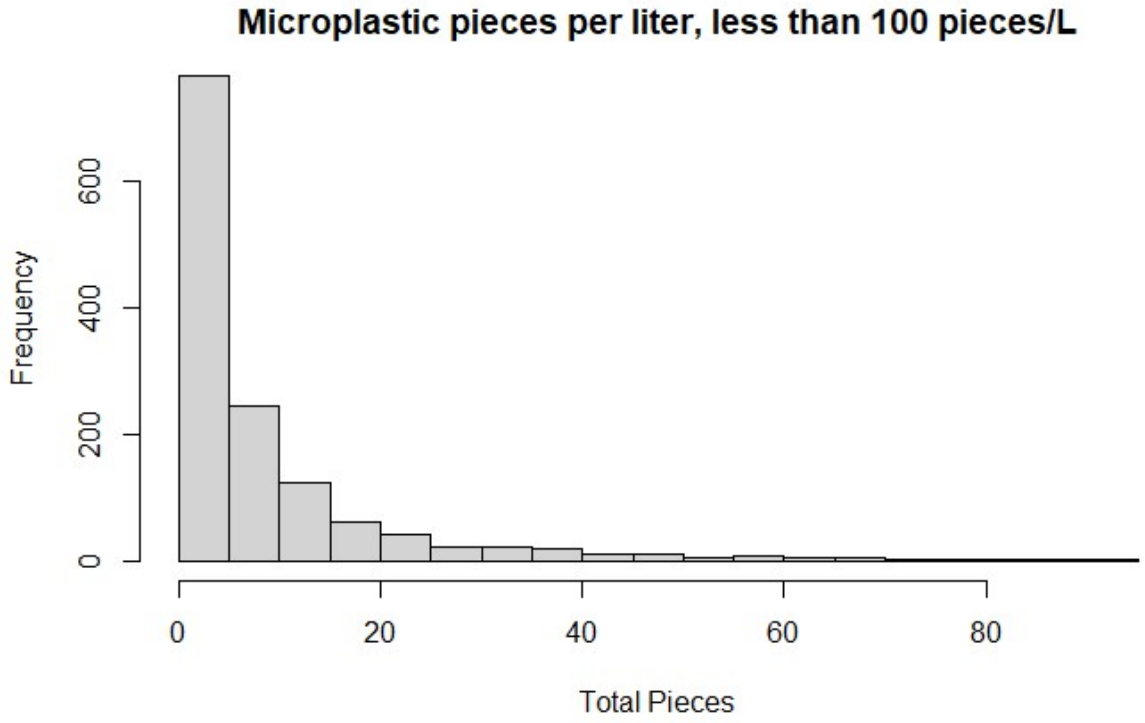



One interesting feature of the dataset is the wide range in the microplastics counts. There are 186 samples where there were no microplastics found in the sample. There are also 131 samples with a count greater than 30 microplastics per sample, and 29 samples with more than 100 pieces per sample. The histogram below shows the microplastic count per liter for all 1393 samples in the dataset. It is extremely skewed right with a minimum value of 0 and a maximum value of 243.



| Summary Statistics, Microplastics per liter | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| 0.000 | 1.538 | 4.167 | 11.801 | 10.909 | 243.077 |

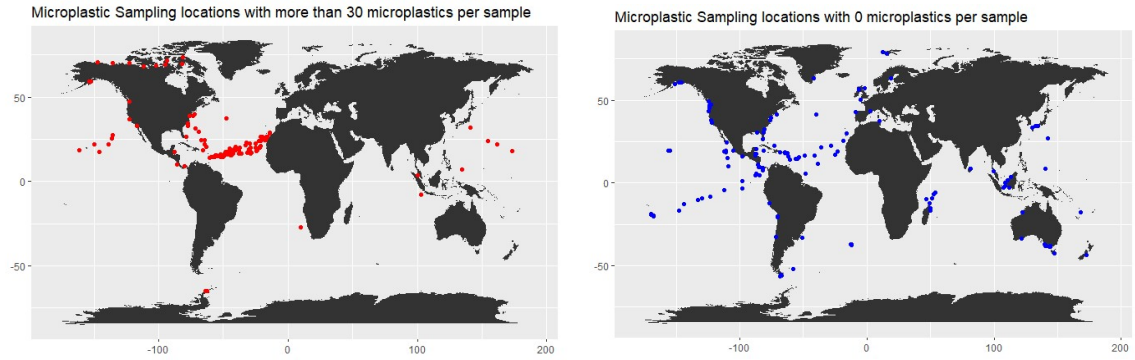
If we exclude the samples with a microplastics per liter rate greater than 100, we can get more of a sense of the shape of the distribution. This is simply "zooming in" to the left side of the distribution.



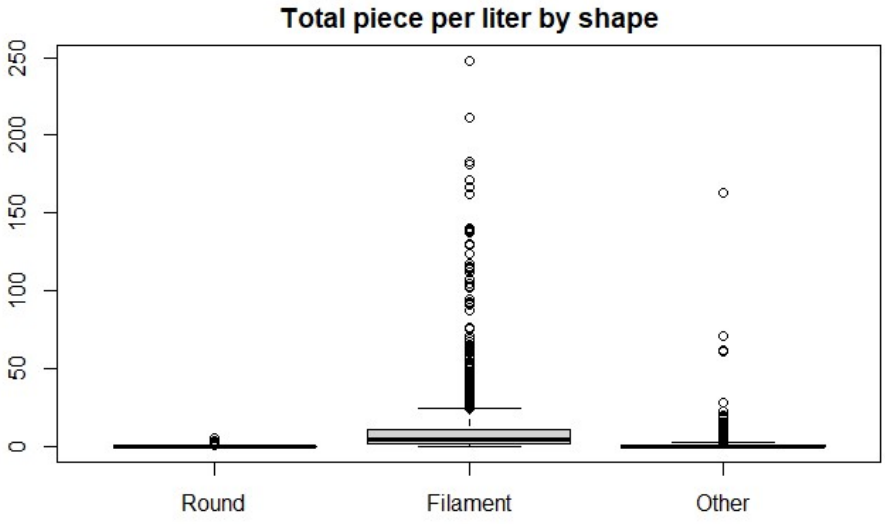

| Summary Statistics, Microplastics per liter (samples with fewer than 100 pieces per liter) | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| 0.000 | 1.538 | 4.000 | 9.043 | 10.400 | 92.308 |

As we begin to model this data, we will have to keep in mind its highly skewed structure, as well as the high number of 0 values. Intuitively, we would want to know both where there are places with very low microplastic counts and places with very high microplastic counts, so it would not be good to ignore either one. Places with microplastic rates of more than 30 pieces per liter are mapped below in

red, and places with microplastic rates of zero are shown in blue. We can see that these two groups overlap spatially in the middle Atlantic. Otherwise, the two groups remain distant from each other and somewhat clustered within themselves.




The three shapes recorded were round, filament, and other shaped plastics. Round shaped plastics would include microbeads, formerly used in beauty products as an exfoliant. Filament shaped plastics would include plastic fibers that are shed from garments when washed. Other shaped plastics includes anything nor in the other two categories. Filaments were present in 1193 samples, other-shapes were found in 460 samples, and rounds were found in only 19 samples, making filaments by far the most common shape. Further, there were 737 samples where filaments were the only shape found. When looking at boxplots of the number of pieces of each type found in the samples, we see that the filaments have a much wider distribution and higher maximum value than either of the two other categories. This conforms with the observations done by Barrows et al. in their 2018 paper based on the same dataset.



Many microfibers, a prominent form of filament, are released when washing clothing made of polyester (Kennedy, 2020). There is currently nothing preventing those fibers from entering waterways and eventually the ocean. This would be an excellent place for policy makers to reduce the microfiber pollution.

## IV. Overview of Spatial Statistics

The method used to analyze the global microplastics distribution is a form of spatial statistics called kriging. In order to understand the methods that follow, it is important to have an overview of this topic. The following section provides that overview in the context of this project.

The purpose of kriging is to estimate the value of some stochastic spatial process $\{\mathrm{Y(s)},\ \mathrm{s}\ \in \mathrm{D}\}$, where $D \subseteq \mathbb{R}^n$ and $n \in \mathbb{Z}$, at an unsampled location s* based on the data from the sampled locations, $\{s_1, s_2,\ \dots, s_k\}, k\ \geq 1, k \in \mathbb{Z}$. In this case, the spatial process $\{\mathrm{Y(s)},\ \mathrm{s}\ \in \mathrm{D}\}$ is defined for $D \subseteq \mathbb{R}^2$ and represents the number of microplastic particles per liter. Arguably, this process should be modeled on $\mathbb{R}^3$ because the Earth is a three-dimensional globe. However, because we will be working with map projections, which flatten the globe onto a conceptual piece of paper, we will be in $\mathbb{R}^2$, where each $s\ \in D$ contains latitude and longitude coordinates.

It is generally assumed in kriging models that the data is Gaussian, meaning that for locations $\{s_1, s_2, \dots, s_k\},\ k \geq 1$, the vector $(Y(s_1), Y(s_2), \dots, Y(s_k))$ follows a multivariate normal distribution. From the histograms in the exploratory data analysis section, we know that our $\mathrm{Y(s)}$, the number of microplastics per liter, does not follow with this.

An important property of spatial data is its stationarity. Generally speaking, this refers to whether the distribution stays the same even if the data is shifted by some distance. Strict stationarity expresses exactly that. For some $h \in \mathbb{R}^n$ and $\{s_1, s_2, \dots, s_k\}\ \in\ D,\ k \geq 1, D \subseteq \mathbb{R}^n$, a spatial process Y(s) is said to be strictly stationary if the vector $(Y(h + s_1), Y(h + s_2), \dots, Y(h + s_k))$ is equivalent in distribution to $(Y(s_1), Y(s_2), \dots, Y(s_k))$. A process is said to be weakly stationary if $\mu(s)\ \equiv\ \mu$, or the mean is the same for all locations, and if $Cov(Y(s_1), Y(s_2))\ =\ C(s_1 - s_2)$, where $C(\cdot)$ is a covariance function (discussed more below). In other words, the covariance between two points can be written as a function based only on location. For many reasons, we would not expect the rate of microplastics in the ocean to be either weakly or strictly stationary, but it is still a good working assumption.

Another important property of spatial data is isotropy. This pertains to the covariance function and its related function the variogram. The variogram is a function related to the variance between the

difference of two values at two locations. When assuming that $\mu(s)$ is a constant and equal to 0, then let

$$var\{Z(s_1) - Z(s_2)\} = 2\,\gamma(s_1 - s_2).$$

The function $2\,\gamma(s_1 - s_2)$ is called the variogram, and $\gamma(s_1 - s_2)$ is called the semivariogram. When we have weak stationarity, we can show the following

$$var\{Z(s_1) - Z(s_2)\} = var\{Z(s_1)\} + var\{Z(s_2)\} - 2\,cov\{Z(s_1), Z(s_2)\}$$

$$var\{Z(s_1) - Z(s_2)\} = cov\{Z(s_1), Z(s_1)\} + cov\{Z(s_2), Z(s_2)\} - 2\,cov\{Z(s_1), Z(s_2)\}$$

$$var\{Z(s_1) - Z(s_2)\} = C(s_1 - s_1) + C(s_2 - s_2) - 2\,C(s_1 - s_2)$$

$$var\{Z(s_1) - Z(s_2)\} = 2\,C(0) - 2\,C(s_1 - s_2)$$

$$2\,\gamma(s_1 - s_2) = 2\,C(0) - 2\,C(s_1 - s_2)$$

$$\gamma(h) = C(0) - C(h)$$

Where $C(\cdot)$ is a covariance function. Now back to isotropy. If $\gamma(h)$ can be written as $\gamma_0(\,||h||\,)$ where $||h||$ is the length of vector h, then the process is isotropic. In other words, a process is isotropic if the semivariogram only depends on the distance between two points. If a process is anisotropic, then the semivariogram can be written as $\gamma(h) = \gamma_0(\,||Ah||\,)$ where A is a $d \times d$ matrix. This matrix A represents a linear transformation of $\mathbb{R}^n$, and when A is the identity matrix, then we are back to the isotropic case. There are methods to handle anisotropic processes, but because we will be assuming that the distribution of microplastics in the ocean is isotropic, we will not go into the details here.

When the variogram is isotropic, $\gamma_0(\,||h||\,)$ can taken many different forms, depending on the shape of the data. Potential functions for $\gamma_0(\,||h||\,)$ include a linear, spherical, Gaussian, Matern, and exponential functions. Each is written as a function of h and some vector of parameters $\theta$. We will now describe the Matern and exponential functions because these are used in the analysis. The Matern function is one of the more mathematically complicated functions dependent on a scale parameter and a shape parameter, both of which must be greater than 0. When the shape parameter is equal to $\frac{1}{2}$, then this corresponds to the exponential form, used in this analysis. The exponential function can be written as

$$\gamma_0(t) = \begin{cases} 0 & if\ t=0 \\ c_0 + c_1(1 - e^{-t/R}) & if\ t>0 \end{cases}$$

Where $c_0$, $c_1$, and R are positive constants. When t is very near to 0, $\gamma_0(t) \sim c_0$. This value is called the nugget. It is similar to a y-intercept, except because it is only defined when t > 0, it never actually touches the y axis. As t increases, the function will increase and then begin to level off. When the function levels off, the value of $\gamma_0(t)$ is called the sill. The sill is attained at a finite value t = R, called the range. Together, the nugget, sill, and range define a classic variogram with a concave shape that levels off as t increases.

Once there is evidence in the data that suggests there is a spatial dependence in the data, and once a covariance function has been selected, then the next step is to estimate the parameters of the covariance function given the data. This can be done with a method of moments estimation, least squares estimation, maximum likelihood or restricted maximum likelihood estimation (REML), or Bayesian estimation. Because the functions used in this analysis use REML methods, we will go into more detail concerning this method.

The maximum likelihood estimator relies on maximizing the likelihood or log-likelihood function. This depends on the underlying distribution of the data, usually assumed to be a multivariate normal distribution. However, the estimator for the variance of this distribution is biased. The restricted maximum likelihood estimator uses a vector of contrasts to find the unbiased estimator of the variance. The likelihood function is still used to find the parameters in vector $\theta$ that maximize the likelihood function. The derivation of how to solve for this is too complex for the purpose of this section, which is to give a general overview of the spatial statistics used in this project.

Once the parameters $\theta$ have been estimated, then the next step is to compute a kriging model. Named for it's originator, kriging uses the existing data and the parameters of the covariance function (also computed given the data) to estimate the value of the spatial process at an unsampled location. Written in another way, given $y(s_1), y(s_2), \ldots, y(s_n)$, we want to predict the value $y(s_0)$ for $s_0 \notin \{s_1, \ldots, s_n\}$. Again, the data is assumed to have a vector mean and matrix covariance. The covariance matrix $\Sigma$ is computed using the covariance function, where $\Sigma_{i,j}$ is the number in the ith row and jth column of $\Sigma$ and $\Sigma_{i,j} = \gamma_0(s_1 - s_2)$. A regression model is then computed, and parameters for the coefficients are estimated. Then, for any new point, the value can be predicted. The result is an optimal linear interpolator for the space.

While there could be many more details added to this section, this should be sufficient to give the reader a sense of the types of analysis done in this report.

## V. Variogram Models

The first step for modeling a spatial distribution is to find a model for the variogram. As mentioned earlier, for this project, the stochastic process of interest is the rate of microplastics per liter of ocean water, referred to as $\{\mathrm{Y}(\mathrm{s}),\ \mathrm{s}\ \in \mathrm{D}\}$ for $D \subseteq \mathbb{R}^2$. The two-dimensional plane is taken to be the pair of latitude and longitude coordinates for each point. While it must be remembered that the Earth is a sphere and all maps are a faulty projection of the surface, the variogram uses the distance between two points. This can be calculated accurately and reliably, regardless of the map projection used to visualize the data. In this case, we use the Haversine method for calculating the distance between two points, otherwise known as the great-circle distance. In this calculation,

$$d(s_1, s_2) = 2r\ arcsin(\sqrt{sin^2(\frac{y_2 - y_1}{2}) + cos(y_1)\ cos(y_2)\ sin^2(\frac{x_2 - x_1}{2})})$$

where $s_i = (x_i, y_i)\ and\ x_i\ is\ the\ longitude\ and\ y_i\ is\ the\ latitude$. This function computes the distance between two points under the assumption that the Earth is a giant sphere. The Earth is actually an ellipsoid, but it is a commonly used method.

Before embarking on calculating a variogram, we would like to ensure that the necessary assumptions hold, or at least are not blatantly broken. Variograms should be stationary and isotropic (though there are tools to handle anisotrophic cases). Stationarity means that the distribution for one group of points has the same distribution as the same group shifted by a random distance. Due to ocean currents and gyres, we do not expect this to hold. For example, the distribution of microplastics in a gyre such as the slow-moving Great Pacific Garbage Patch may not be equivalent in distribution to a sample along the fast-moving Atlantic Gulf Stream. However, we will assume that the distribution is stationary. If a process is isotropic, then the variogram only depends on the distance between points. There is no geometric transformation needed in isotropic cases. We assume that the process is isotropic because the rate of microplastics per liter at a location should be directly related to other locations by the distance. Finally, we will assume that the covariance function is positive definite.

Before attempting to answer whether the data from 2013 through 2017 should be treated as one unit of broken down into their individual years, we needed to find or create code to calculate our variogram. Two potential candidates were the *variog()* function in the package *geoR* and the *fit.variogram()* function in the *gstat* package. These ultimately did not prove to be very useful. The former resulted in a plot that did not conform to the classic variogram shape and that could not be

easily adapted to change the distances and tolerance region sizes while the latter required a very specific data format that was more difficult to figure out than necessary.

In order to address this, the code was written by hand by the author. The code is included in the Appendix of this paper. The function used the Haversine method for calculating distance between two points and recorded the distance in kilometers. Because the data was not collected on a grid, a changeable tolerance region was included in the function. Further, the starting and ending distance were changeable, meaning that variogram could start at 10 kilometers and go through 500 kilometers or start at 100 kilometers and go through 300 kilometers, as desired by the user. A method of moments calculation was used.

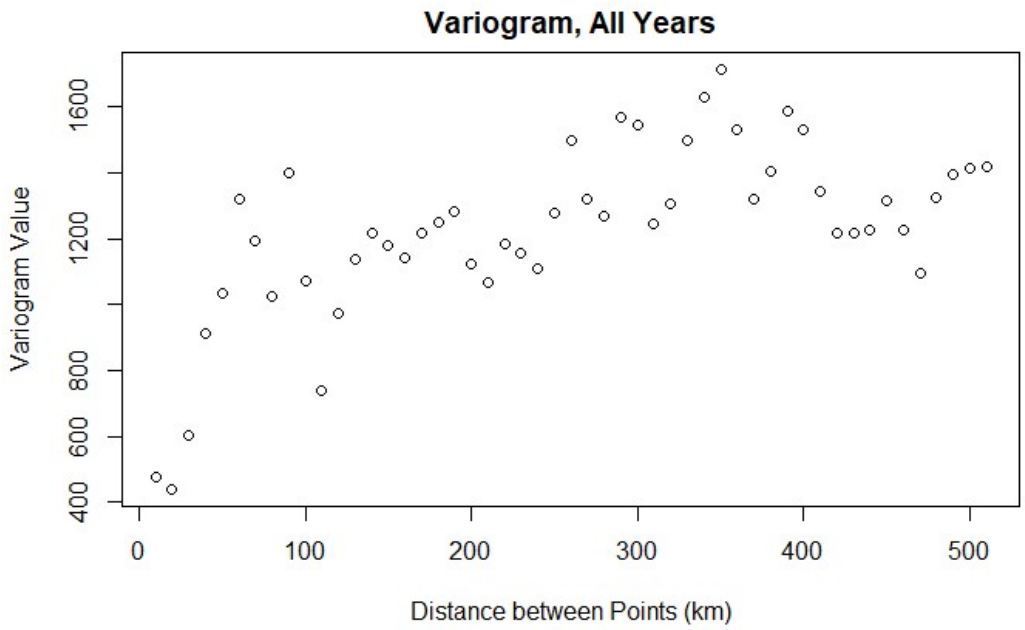
The following variogram was calculated from 10 kilometers to 510 kilometers with a tolerance region or buffer of 10 kilometers. Because the data was not on a standardized grid, a tolerance or buffer region was used so that datapoints would actually be included in the estimate. This means that for the method of moments variogram estimator,

$$2\,\hat{\gamma}(h) \;=\; \frac{1}{|N(h)|} \sum_{(s_i, s_j) \in N(h)} \{Z(s_i) \;-\; Z(s_j)\}\ \text{^2}$$

Where N(h) refers to the set of points within some small neighborhood around h. In mathematical notation,

$$N(h) \;=\; \{(s_i, s_j) : \; s_i \;-\; s_j \in \; T(h)\,\}$$

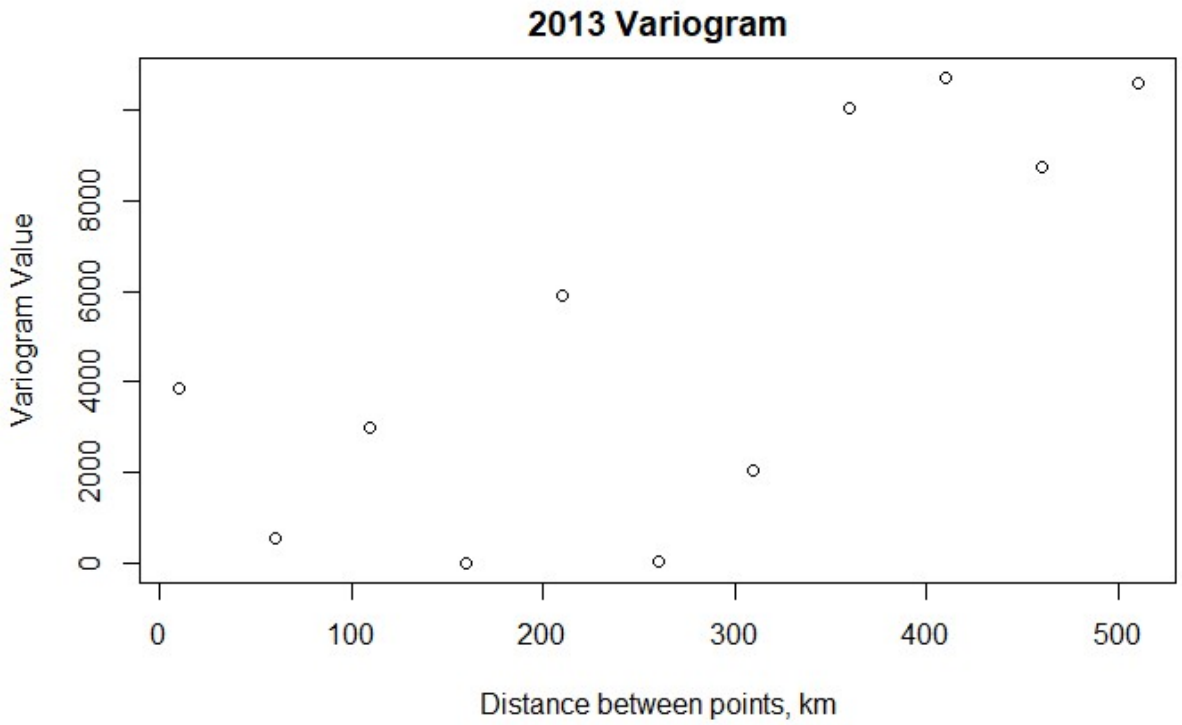
Where T(h) is the small neighborhood around h. This way of plotting a variogram does give promising evidence of spatial correlation because we are seeing the classic concave curve shape with a nugget, sill, and range.



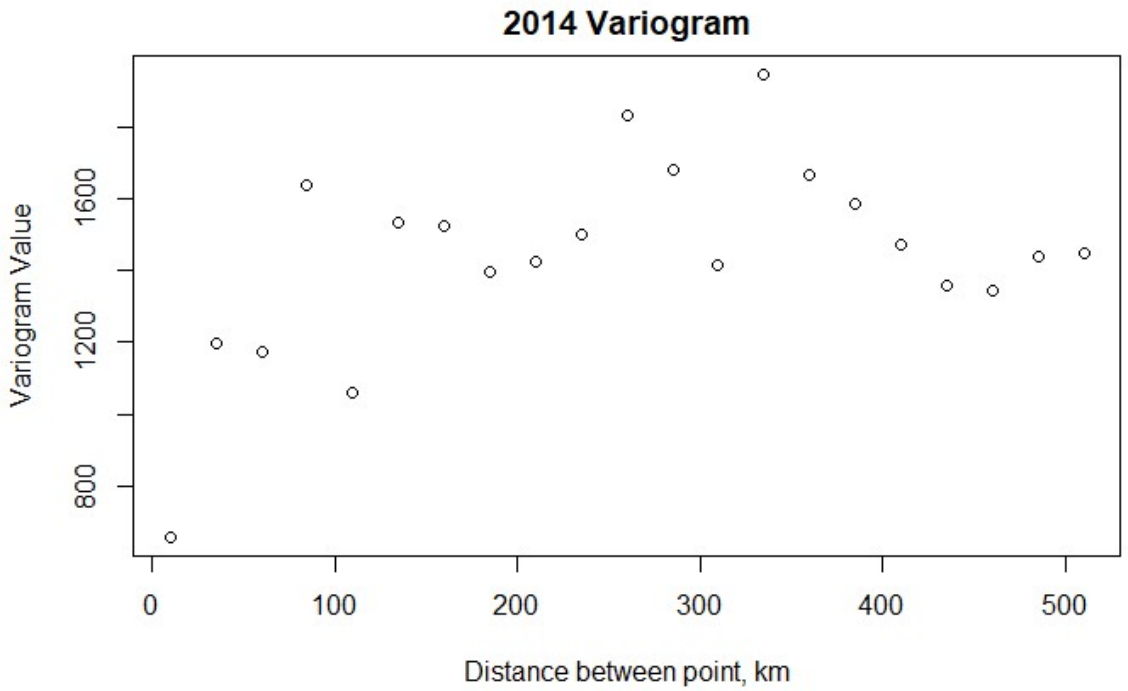

## Variograms by Year

In order to address the question of whether or not a temporal component should be included in the model, variograms were calculated for each year of data from 2013 to 2017. The number of points in each year varied widely, with 50 samples in 2013, 725 samples in 2014, 318 samples in 2015, 249 samples in 2016, and 51 samples in 2017. Maps of the point locations for each year are provided in the appendix. The yearly variograms tended to also show evidence of spatial correlation through their variogram shape, but the years with fewer points showed this less clearly. This could be due to the low number of points or to the way the data was binned, but it also sparked the idea for attempting kernel smoothing, as discussed later in this report. When viewing the map of the sampling locations for 2017, there is a clear group of points off the southern tip of South America and another off the northern coast of Madagascar. These two places are quite far away from each other and may experience different means and variances in their microplastic distribution. More on this will be discussed in the Kernel Smoothing section.

The variogram for the 2013 data was constructed with 50 samples over a distance of 10 to 510 kilometers in 50-kilometer increments. In this case, the tolerance region was set to be a 25-kilometer radius circle around h. The plot does not show the classic variogram shape but does show an overall increase in the variogram value as the distance between points increases.



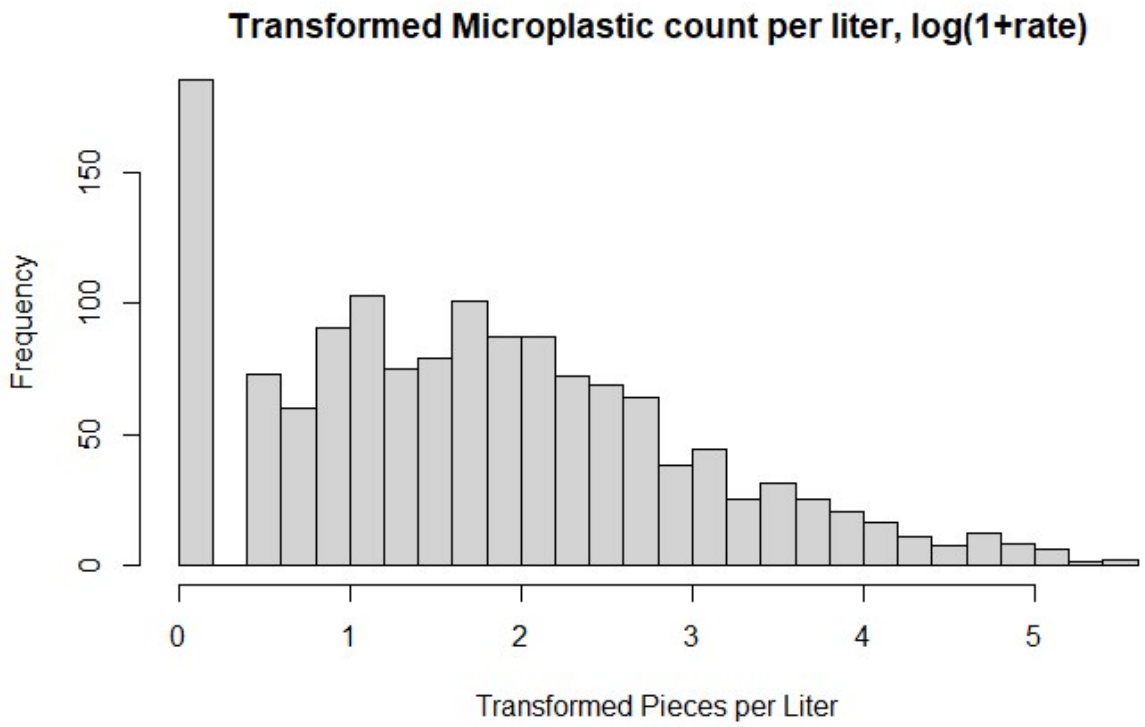

The variogram for the 2014 data was constructed with 725 samples over a distance of 10 to 510 kilometers in 25-kilometer increments. The tolerance region was set to 10 kilometers. The plot shows evidence for spatial correlation.



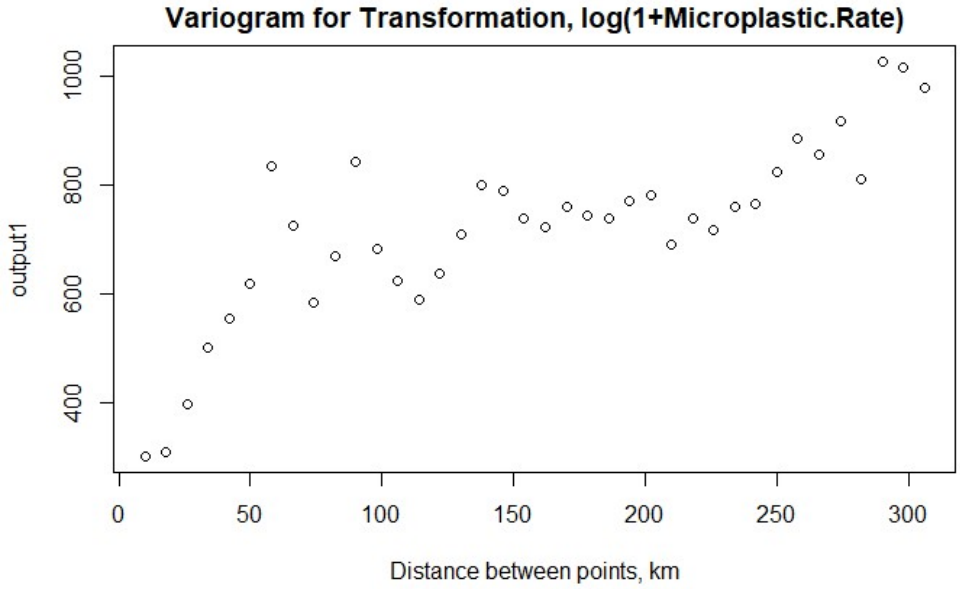

The variograms for the other years looked similar to the two shown here and were not included for the sake of brevity. There is some evidence that each of the years show spatial correlation, and it may be interesting to add a temporal component to the analysis. In this analysis, a comparison will be done between data collected before 2015 and data collected after (and including) 2015.

## Variogram for Transformed Data

Because of the highly skewed nature of the data, several transformations of the microplastic count per liter were investigated to try and make the data more Gaussian in nature. The square root of the microplastics per liter, a logarithmic transformation, and a kernel smoothing method designed to standardize means and variances across different regions. The kernel smoothing method results were not as expected and are discussed later, but the most promising of these transformations was a logarithmic transformation, $\log(1 + microplastics.per.liter)$. This resulted in the following histogram:




This plot remains skewed but less drastically. The range of values is much more condensed and manageable than the untransformed data. The large bar at the zero represents samples that had no microplastics in them. The rest of the plot appears almost like a Gaussian distribution or a Poisson distribution. Because of this, models for a truncated Gaussian or zero-inflated Poisson model will be considered. The variogram of the transformed data also provides evidence of spatial correlation.



The remainder of the modeling is based off of this transformation. So while the spatial process of interest is the number of microplastics per liter $\{Y(s)\}$, the process being modeled is $\{Z(s), s \in D, where\ Z(s) = log\big(1 + Y(s)\big)\}$.

## VI. Restricted Maximum Likelihood Estimators for Spatial Parameters

After seeing that there was evidence for spatial correlation in both the untransformed and transformed data across all years, the next step was to find estimates for the covariance function parameters. This was done using the function *spatialProcess()* from the R package *fields*. When given a specific covariance function from, the *spatialProcess()* function uses the restricted maximum likelihood estimator to estimate the nugget, sill, and range, as mentioned earlier in the spatial statistics overview section. The function needed only unique values in order to work, and there were 27 data points in the dataset that were considered duplicates by the *spatialProcess()* function. These points had the same latitude and longitude values, but different microplastic counts. Interestingly, each pair of duplicate points was also taken on the same date. It is possible that the samplers took multiple samples from the same general site and entered the same location for these samples. Regardless, these duplicates were subsetted out for the running of the function, with the duplicated entry being removed chosen at random, leaving a dataset with 1366 values.

The *spatialProcess()* function estimates the nugget $\sigma^2$, process variance $\rho$, and range $\theta$ parameters of the covariance function. The sill of the function is equivalent to $\sigma^2 + \rho$. It uses the REML method to estimate these parameters. An exponential model was used for the initial variogram estimation of the log transformation $Z(s) = log\big(1 + Y(s)\big)$ with the following results.

| | Nugget | Sill | Range (miles) |
|---|---|---|---|
| Value, *spatialProcess()* | 0.9156 | 1.8822 | 12.8644 |

While other models were used earlier in the process (Matern with smoothness 1, Wendland model), it was decided that the exponential would serve as a stable and very similar model to the other options.

An alternative method for estimating the parameters was attempted as well. The function *likfit()* from the package *geoR* was also used for estimating the variogram parameters. It also used Restricted Maximum Likelihood estimators, and returned the following results:

| | Nugget | Sill | Range (miles) |
|---|---|---|---|
| Value, *likfit()* | 0.8402 | 1.8515 | 13.8716 |

These results are similar, but not the same. This comparison was done because *geoR* functions were used for the kriging step, so it was important to verify that *fields* and *geoR* would return similar results. While the results are not equal, they are of the same magnitude, so the analysis was continued.

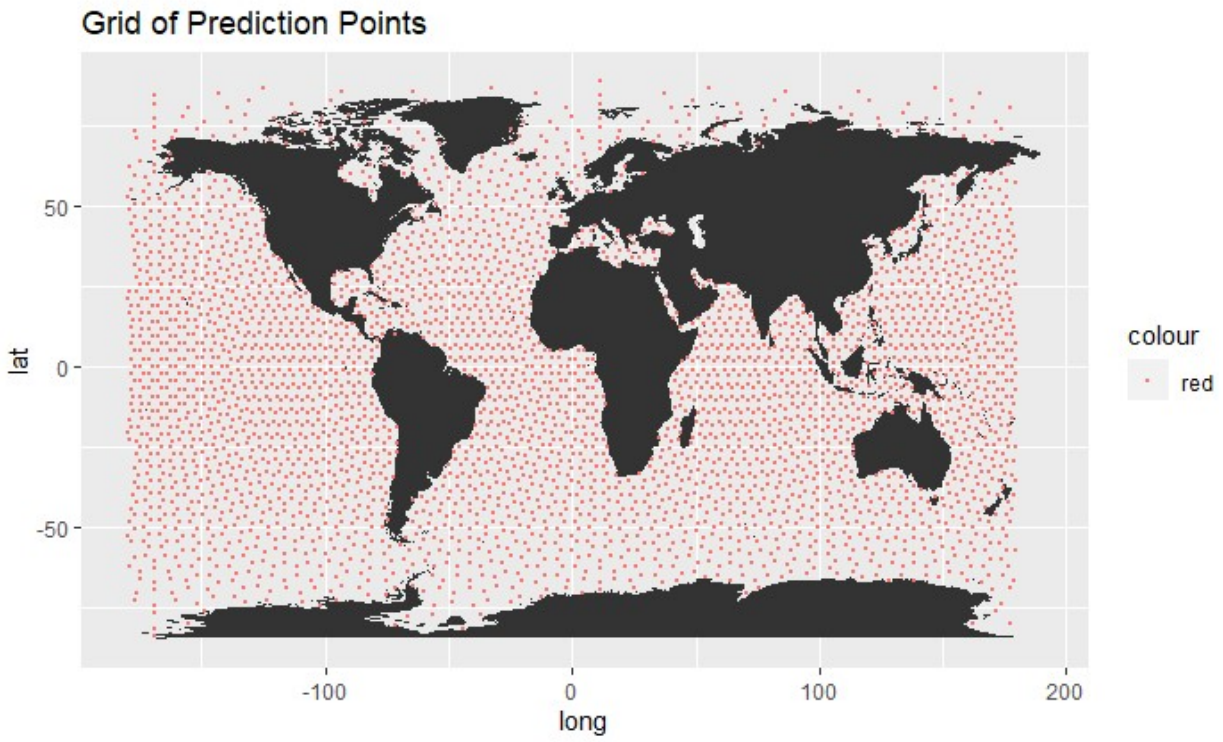
It is worth taking a moment to consider whether the REML method is the most viable to use for this data. As we know, the REML method assumes the data is Gaussian, an assumption that does not hold in this case. The variogram could have been fitted with more crude methods instead. However, this does not mean that we cannot use the REML method, especially because the trouble with this dataset is the high rate of 0's in the data, signifying no microplastics in the sample. REML and kriging are both used in cases where the Gaussian assumption may not hold, so it is worthwhile to explore what can be done with these Gaussian models. Nevertheless, these assumptions may be broken to a sufficient degree that an alternate, likely Bayesian, approach may have to be taken. This Bayesian approach is briefly discussed towards the end of this report.

## VII. Kriging Models

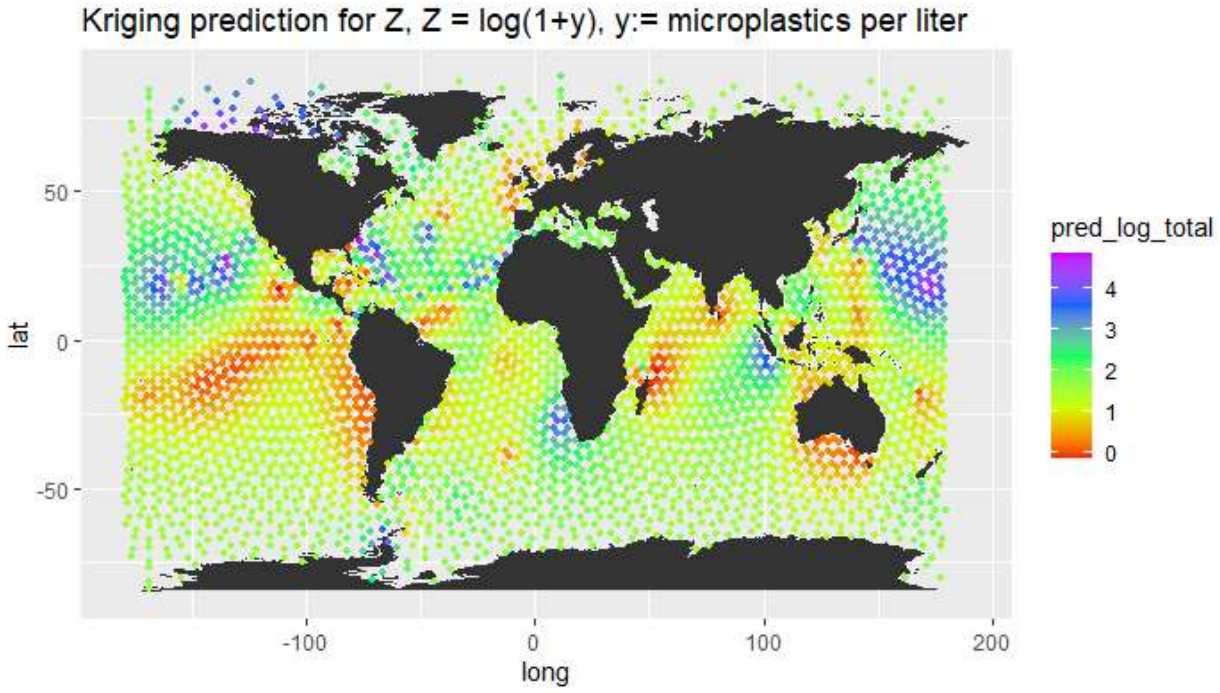
The next step for modeling the spatial distribution of ocean microplastics is to use a kriging method. This process uses the data provided and the covariance model parameters to predict the value of the microplastics at all unsampled locations in a region. It assumes that the data in question is Gaussian, an assumption that does not, unfortunately, seem to be upheld in this case. Nevertheless, we will still attempt a preliminary kriging model.

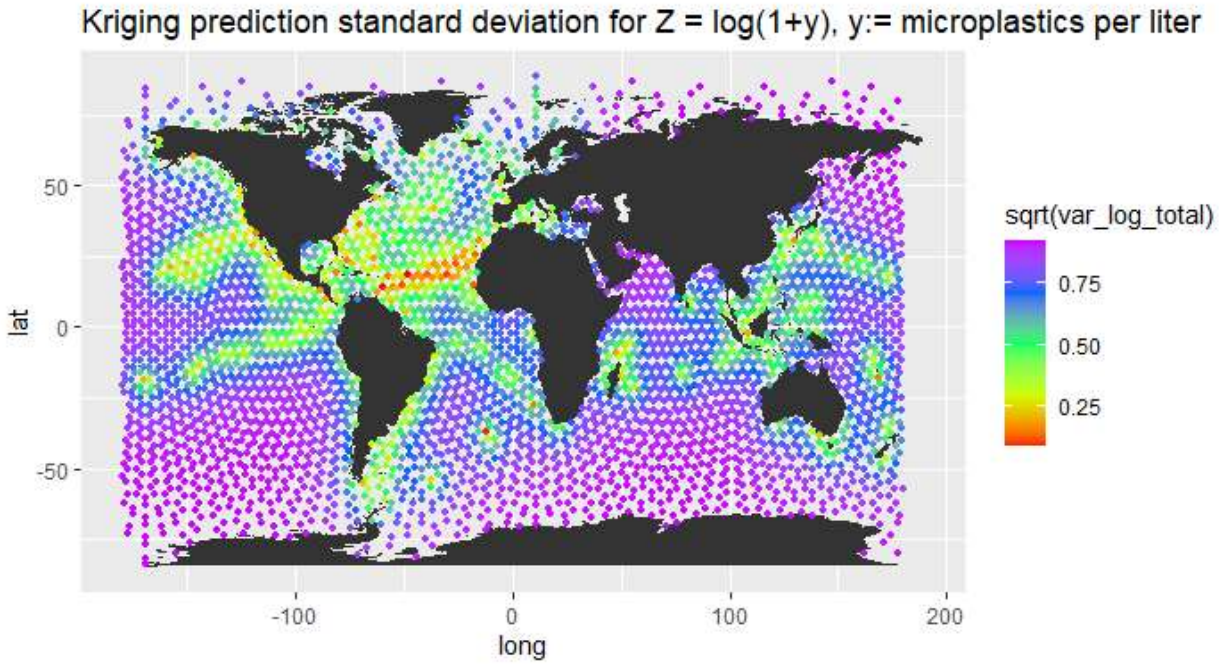
The R code package *geoR* was used for kriging, and the function used was *krige.conv()* This was used because of the usability factors. It was easy to predict points on a grid with this function, to map them, and to then retrieve those values for analysis. The equivalent *fields* functions, *Krig()* and *surface.Krig()* were not as usable on those three facets.

The model was predicted on a hexagonal grid of points covering the globe. This was done using the packages *devtools*, *dggridR*, and *rgdal*. The initial output grid included points on both land and ocean. Because microplastic estimates we not wanted on landmasses, points over the land were filtered out using the GIS software ArcGIS. The resulting grid contained 11,900 points and is shown below.




When the kriging function was run, the following graph was produced. Blues and greens indicate higher log-microplastics rates while reds and oranges indicate lower log-microplastic rates. The variance plot is also included.

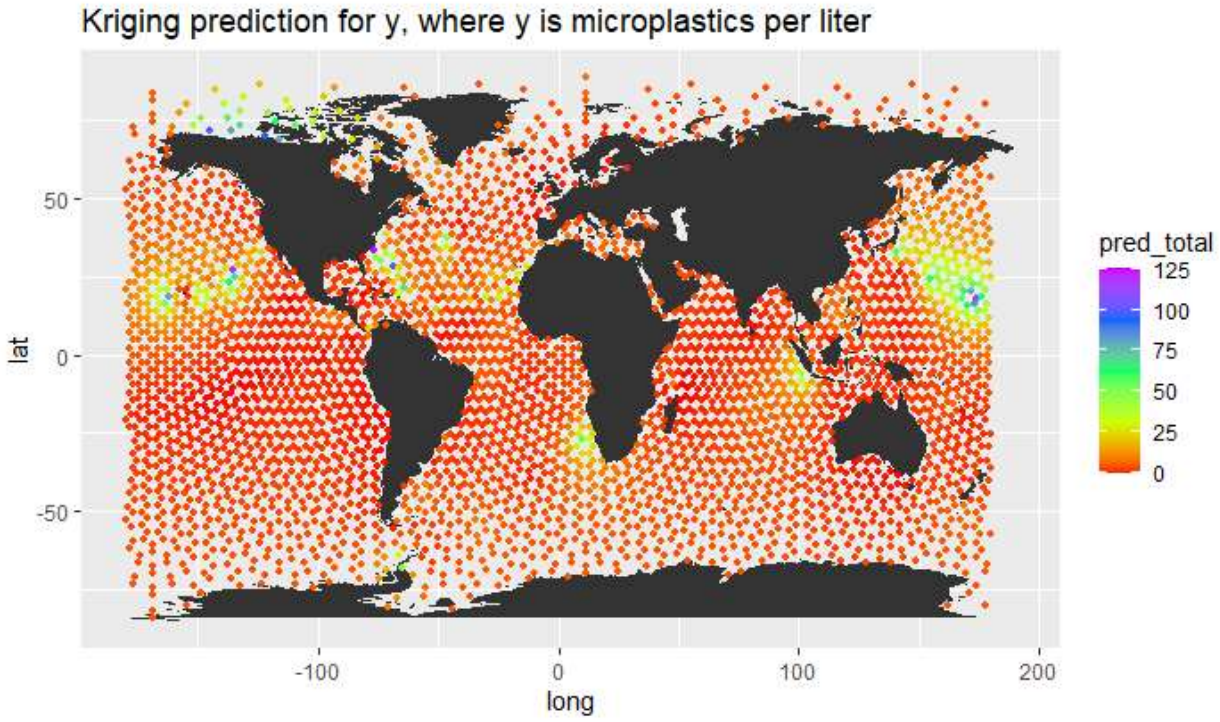



From this, we see that there are microplastic hot-spots in the Arctic Ocean above North America, to the east of Hawaii in what is known as the Great Pacific Garbage Patch, in the Caribbean, the southwest

coast of Africa, and southwest of Sumatra. These results are somewhat expected (Barrows, 2018), and a more more specific comparison is done later. There are microplastic cold-spots along the Pacific coast of South America, waters surrounding Britain and Scandanavia, and the south of Australia.

When the data was transformed back to its original scale, the resulting map was much more difficult to interpret visually. It is included below.




Models were also calculated that exclusively examined the distribution by shape (filament, round, other shape) and by color (black, blue, red, transparent, other color). This was done for all years combined, also on a log scale. Each model was given a subset of the data corresponding to the variable of interest, and variogram parameter estimation, kriging model, and map was done. For the sake of brevity, these results are not included here. They can be found in the appendix.

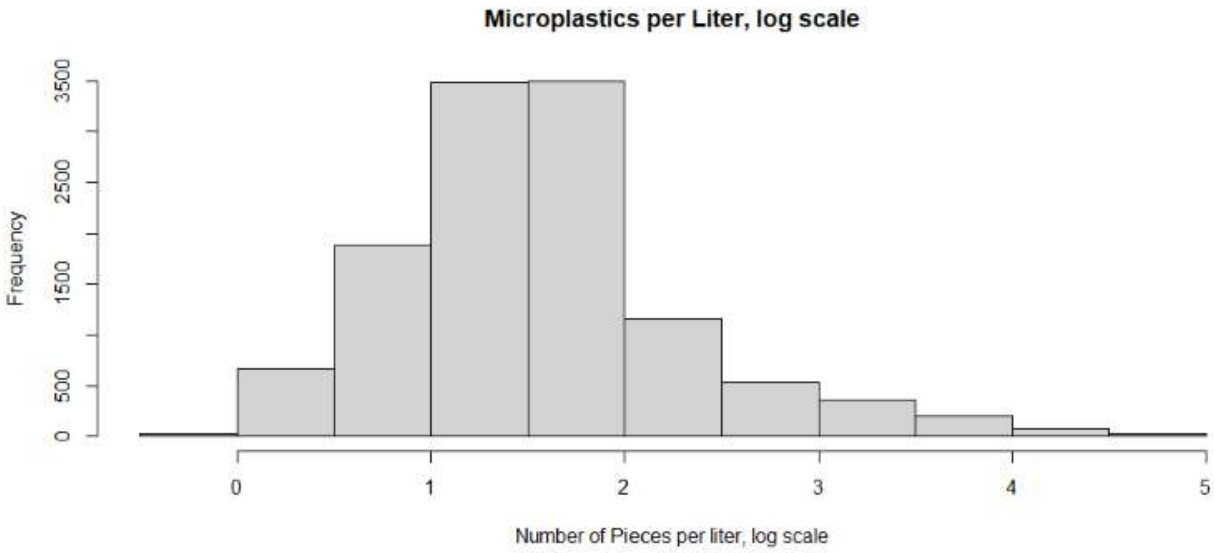
Models were then calculated in order to assess whether there was any time effect present in the data. The data was grouped into a before-2015 subset (including data from 2013 and 2014) and an after-2015 subset (including data from 2015, 2016, and 2017). This division had a similar magnitude of points in each group, 759 and 607 respectively. The process was repeated for log- micropplastic-rates on all plastic types combined, as well as by color and by shape. It was anticipated that the later time group

would have more microplastics per liter than the earlier time group. However, the data does not show this.

# VIII. Analysis of Kriging Results

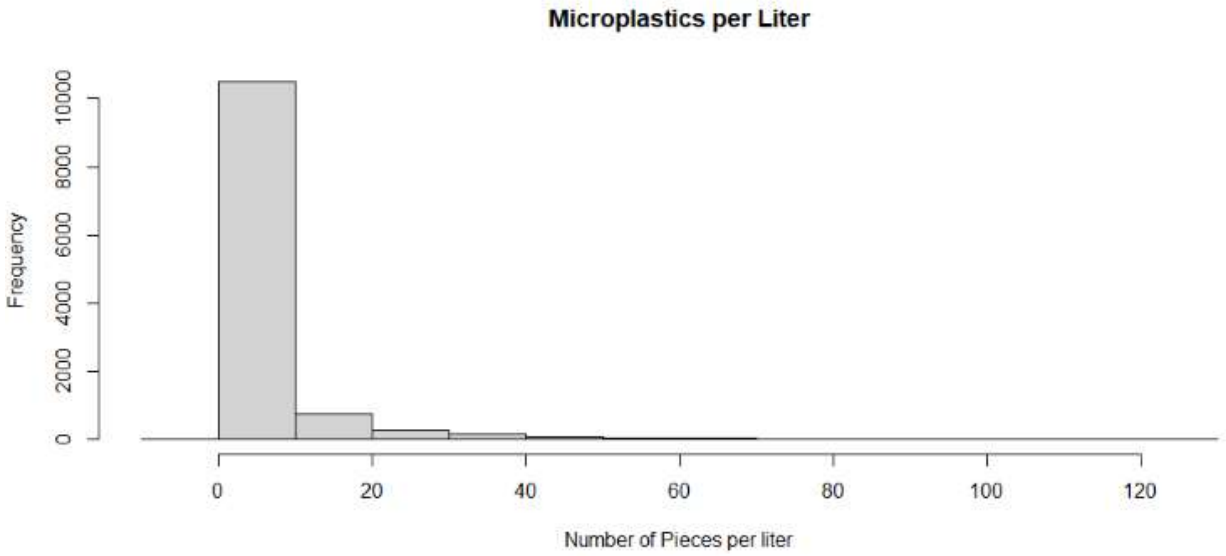
## Overall

We will now look at the results for the microplastic rates over all years and all variables combined. There were 11900 data points predicted.




| Summary Statistics, Microplastics per liter prediction, on a log scale | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| -0.1238 | 1.0690 | 1.4871 | 1.5429 | 1.8442 | 4.8407 |

The variance was 0.5590. The negative values, which are, of course, unrealistic, are a result of the high number of 0 entries in the data coupled with a Gaussian assumption.

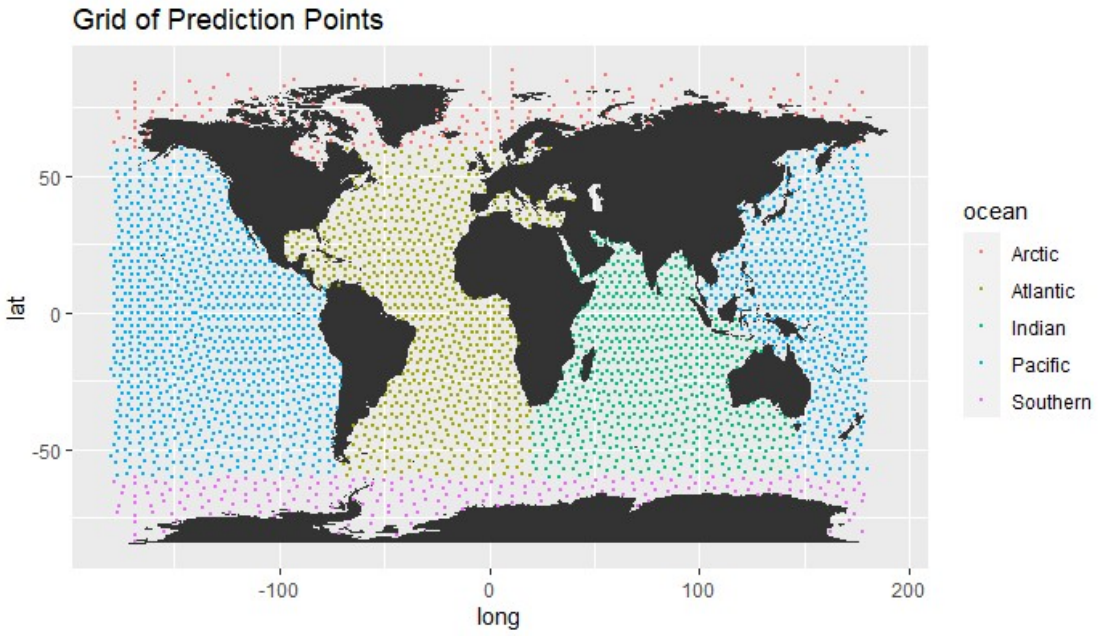



| Summary Statistics, Microplastics per liter prediction | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| -0.1165 | 1.9124 | 3.4241 | 5.7457 | 5.3227 | 125.5628 |

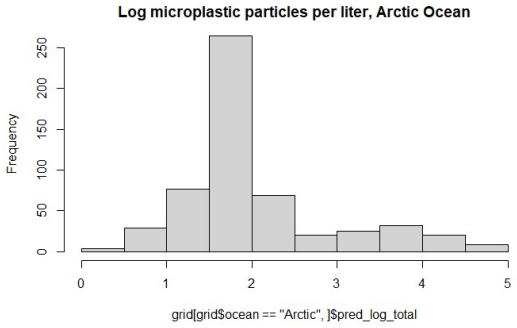
The variance was 86.1444.

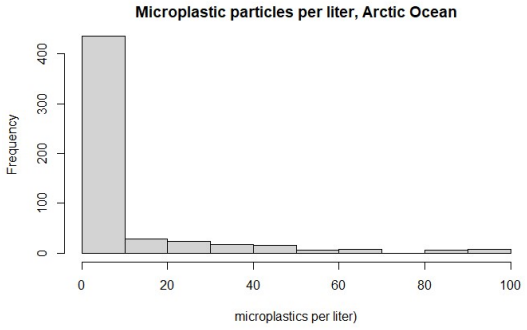
This is a highly skewed distribution, reflective of the input data. There are 21 points that were predicted to be negative on both a log scale and the original scale. It is not known why this happened. The average number of microplastics per liter in the ocean is 5.7457 pieces per liter.

These results were analyzed by ocean. The ocean divisions can be seen below.

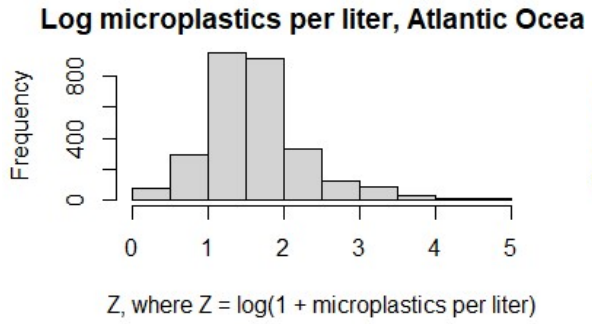


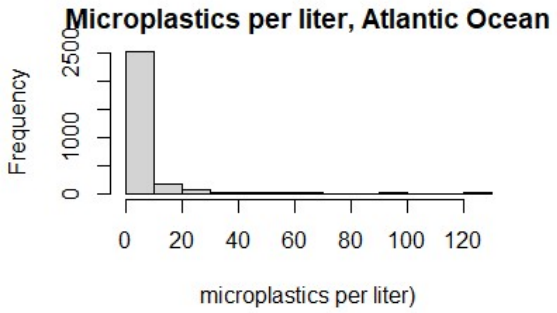

There were 549 prediction points in the Arctic, 2815 in the Atlantic, 2367 in the Indian, 5468 in the Pacific, and 701 in the Southern Ocean.



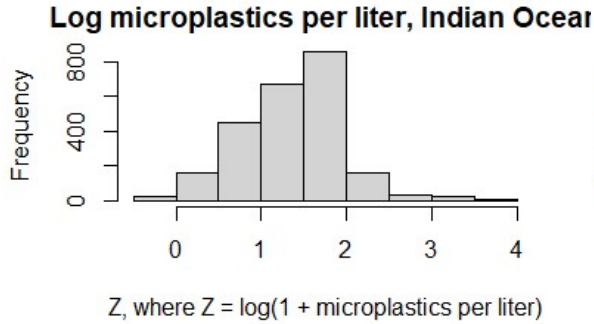



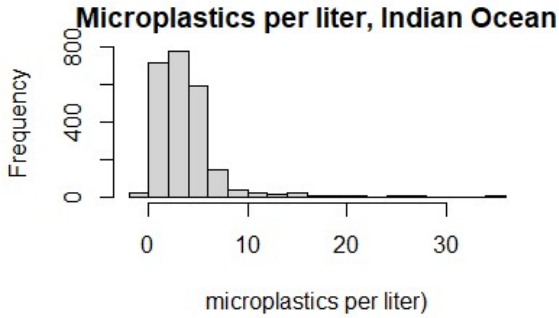

| Summary Statistics, Microplastics per liter prediction for the Arctic Ocean | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| 0.3128 | 3.8858 | 4.5879 | 11.5715 | 7.5007 | 99.8980 |



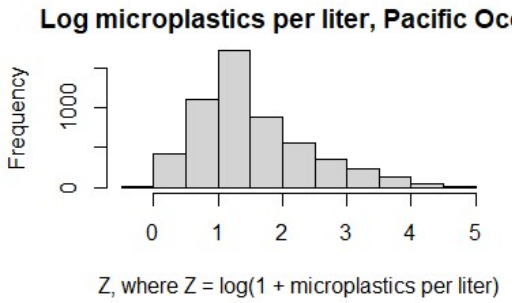



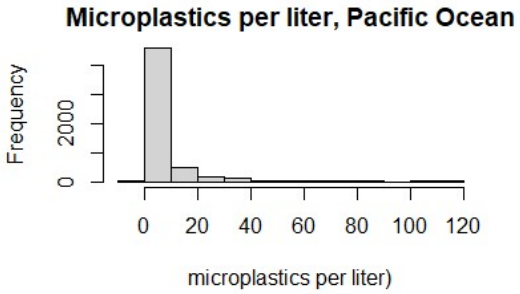

| Summary Statistics, Microplastics per liter prediction for the Atlantic Ocean | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| 0.0128 | 2.2948 | 3.7800 | 5.7018 | 5.6855 | 125.5628 |



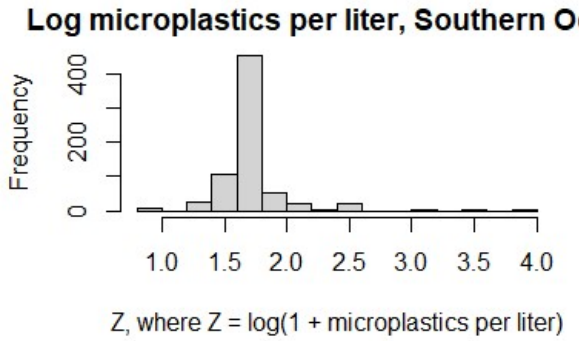



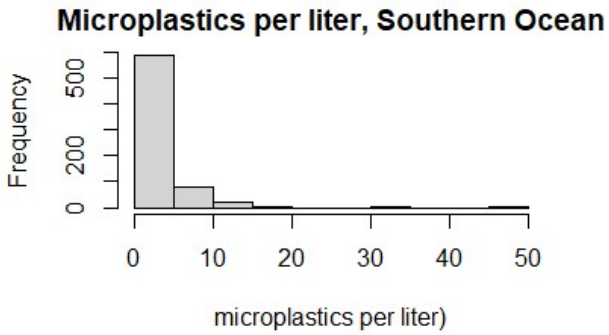

| Summary Statistics, Microplastics per liter prediction for the Indian Ocean | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| -0.1012 | 1.6347 | 3.1336 | 3.6185 | 4.2965 | 35.8395 |




| Summary Statistics, Microplastics per liter prediction for the Pacific Ocean | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| -0.1165 | 1.5753 | 2.8516 | 6.1917 | 6.1373 | 119.2590 |



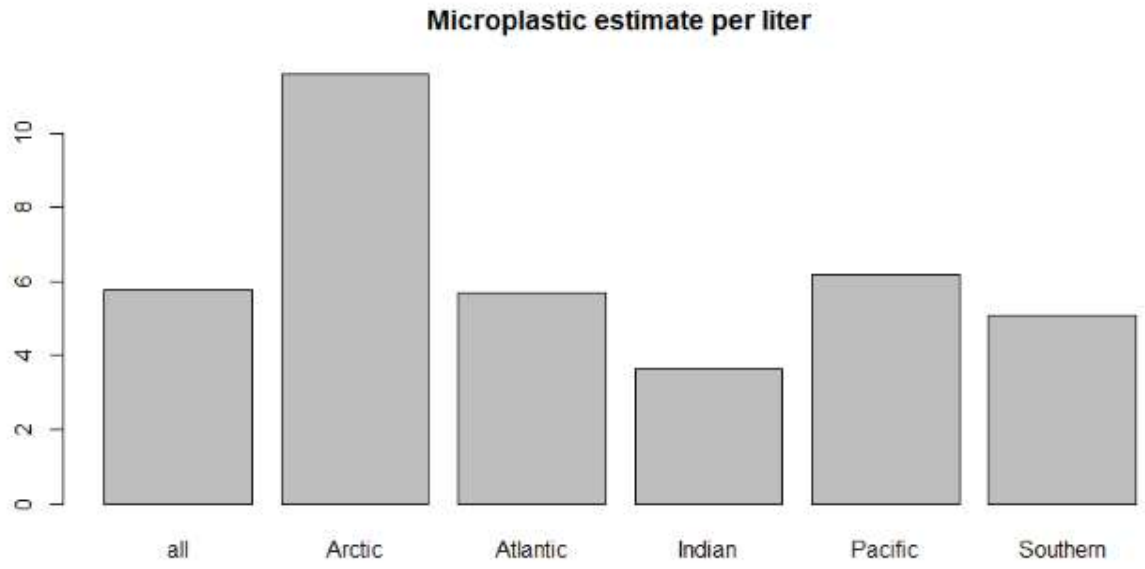




| Summary Statistics, Microplastics per liter prediction for the Southern Ocean | | | | | |
|---|---|---|---|---|---|
| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
| 1.399 | 3.979 | 4.245 | 5.063 | 4.771 | 48.629 |

The following chart displays the mean microplastics per liter prediction for all of the oceans.

| | Average estimate, microplastics per liter |
|---|---|
| All | 5.4757 |
| Arctic | 11.5715 |
| Atlantic | 5.7018 |
| Indian | 3.6185 |
| Pacific | 6.1917 |

| | |
|---|---|
| Southern | 5.0633 |




We see that the Arctic has the highest average prediction value while the Indian Ocean has the lowest prediction value. However, this ocean was not sampled as thoroughly as they could have been, so these estimates come with a grain of salt.

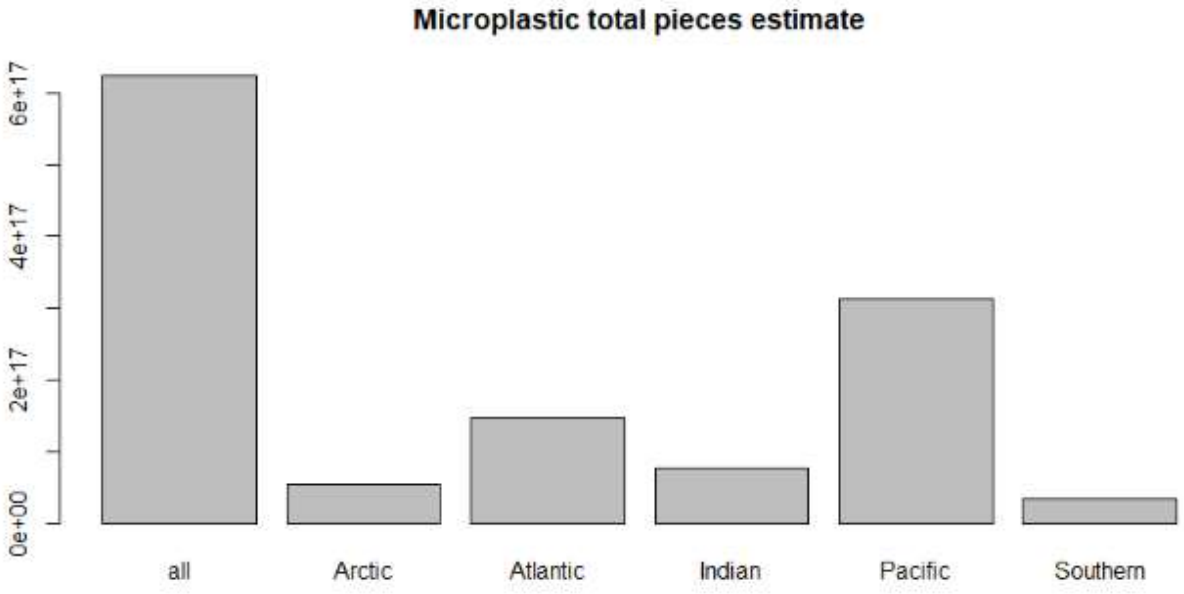
Next, an estimate of the total number of microplastic pieces in the surface of the ocean was calculated. This was done by assuming that samples were coming from the top 30 centimeter of ocean water and by using surface area estimates provided by NOAA. The volume of the surface of the ocean was then calculated as follows:

$$Volume_{liters} = Surface\ Area_{km\text{^}2} * \frac{1000 \times 100cm}{1km} * \frac{1000 \times 100cm}{1km} * 30cm * \frac{1\ liter}{1000cm\text{^}3}$$

The following table shows the steps taken to calculate the estimated number of microplastic pieces in the ocean.

| | Microplastic estimate per liter | Surface Area (kilometers squared) | Volume (Liters) | Piece Count Estimate |
|---|---|---|---|---|
| All | 5.7457 | 361,900,000 | 1.0857e+17 | 6.238149e+17 |
| Arctic | 11.5715 | 15,558,000 | 4.66740e+15 | 5.400885e+16 |
| Atlantic | 5.7018 | 85,133,000 | 2.55399e+16 | 1.456244e+17 |

| | | | | |
|---|---|---|---|---|
| Indian | 3.6185 | 70,560,000 | 2.11680e+16 | 7.659726e+16 |
| Pacific | 6.1917 | 168,723,000 | 5.06169e+16 | 3.134064e+17 |
| Southern | 5.0633 | 21,960,000 | 6.58800e+15 | 3.335705e+16 |




These estimates for the averages were not done with weighted averages based on the variance of the spatial model; meaning, estimates near many original data points are weighted the same as estimates far, far away from any original data points. This is not the best way to handle this because many estimates were generated far from data points, making them less credible.

## By Variable Type

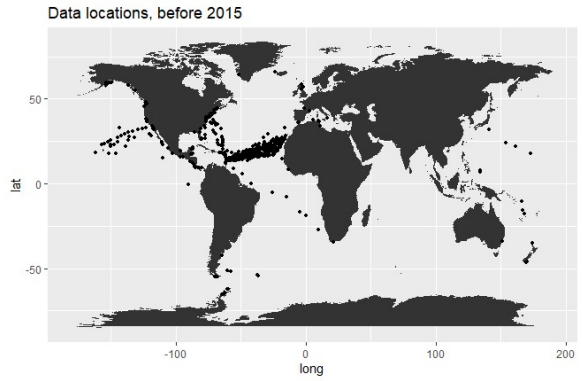
As mentioned earlier, the variogram estimate and kriging model were done separately for subsets of the data. These subsets were based off of the shape and color of the microplastic pieces. While global distribution maps were generated for each subset, they did not differ drastically from the overall distribution pattern; the same hotspots and cold-spots were present. The distributions themselves, ignoring the spatial component do differ somewhat and are included in the appendix. The average and variance of these models are below, calculated in the same way as above.

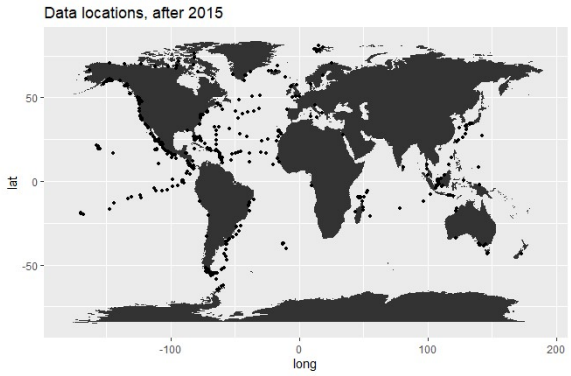
| | Estimate |
|---|---|
| Total | 5.7457 |
| Filament | 5.4901 |
| Round | 0.0115 |
| Other Shape | 0.3004 |
| Black | 0.8996 |
| Blue | 2.0125 |
| Red | 0.6751 |
| Transparent | 2.3487 |

Transparent plastics are the most prominent color, and filaments are the most prominent shape.

## By Year

The models were also done on two divisions of years. The first group of years was 2013 and 2014, referred to as pre2015, which included 759 data points. The second group of years was 2015, 2016, and 2017, referred to as post2015, which included 607 data points. The location of the points can be seen below. As is obvious below, the points do not overlap very much, except in the Atlantic Ocean. It would be better for the comparison if the sampling locations in the different time groups were close to each other, but such is the data.




A summary of the results of the models can be found below. The results are reporting the microplastics-per-liter on an untransformed scale.

| | Pre2015 estimate | Post2015 estimate |
|---|---|---|
| Total | 12.1367 | 3.131 |
| Filament | 11.704 | 3.0692 |
| Round | 0.0181 | -0.00015 |
| Other shape | 0.6509 | 0.1361 |
| Black | 2.0147 | 0.4425 |
| Blue | 2.27 | 1.176 |
| Red | 0.955 | 0.4429 |
| Transparent | 4.9259 | 1.1788 |

From this, we see that the years 2013 and 2014 have higher microplastics rates than the years 2015, 2016, and 2017 across all subsets of the data. This is not what we expected to see and may be due to the sampling locations and values rather than an overall reduction of microplastics in the surface of the ocean, but there is no reason to think these estimates are incorrect. In order to investigate this further, an analysis of point-pairings could be done, with one point in the pre-2015 group and the other located geographically near and from the post-2015 group. However, this would rely on an assumption that the ocean surface is relatively static, which is not the case.

# IX. Discussion

## Model Areas for Improvement

From the very beginning, it was known that the kriging method would not be the best method for modeling the distribution of microplastics in the ocean. This is because kriging relies on an assumption of Gaussian data while the ocean plastic data was heavily skewed right. Thus, all of the estimates are likely overestimates. Further, kriging works best when data points are on or close to a grid, or at least covering the area of interest. This most certainly was not the case, with large swathes of every ocean unsampled. These two properties of the data, its skewness and its sampling locations, do not lend themselves well to this sort of modeling and throw some doubt into the estimates.

When a covariance function is modeling using covariates, the kriging model must also include covariates. This poses an issue for this dataset because of the unknown covariate values for unsampled locations. For example, ocean salinity could be used as a covariate for the covariance function, but because the salinity of every unsampled location in the ocean is not known and available to be used, salinity cannot be used in the model. It would be possible to include a certain indicator variable for the model, one indicating whether a sample was taken on the coast or in the open ocean.

This is a variable included in the Adventure Scientists dataset, and one that could be implemented for prediction sites using geographic information systems techniques to indicate whether or not a point is within 12 nautical miles of a coastline.

### Comparison with Previous Results

The paper by Barrows, Cathey, and Peterson in 2018 is the only other paper found by the writer that addresses works with the same dataset used here. Barrows is a member of the Adventure Scientists team, and the paper gave an overview of their sampling methods and an analysis of the data. In this analysis of the data, the spatial distribution of plastics is ignored, and the data are analyzed as if they had no spatial component at all. With this in mind, Barrows calculated the average microplastic particles per liter as a simple arithmetic mean, where $y_i$ is the rate of microplastics per liter at the *ith* datapoint:

$$\frac{1}{1393} \times \sum_{i=1}^{1393} y_i \; = \; 11.8 \, particles/liter$$

The estimate also included a standard error or 0.6 particles per liter, also calculated with no regard to the spatial component. This is much larger than the estimate of 5.4757 particles/liter obtained through this spatial analysis.

Barrows includes estimates for the Arctic, Atlantic, Pacific, Indian, and Southern oceans as well. In her analysis, the Arctic and the Southern oceans contained the highest average rate of particles per liter, as 31.3 and 15.4 particles per liter, respectively. This analysis ranked the Arctic Ocean as having the highest rate, at 11.6 particles per liter, while the Pacific Ocean was the second highest, with the rate being 6.1917. Again, the Barrows estimates are higher than the estimates obtained here. This may be due to the large amount of 0-valued data entries in the data causing many prediction points to be estimated low. It may also be because prediction points that are far from any sample points are estimated very low, causing the average to drop as well. In reality, there are probably microplastics in the waters predicted to have 0 microplastics per liter in them, so our estimate is probably low. Nevertheless, including the spatial component is very important for understanding the distribution of microplastics in the ocean, and it is worthwhile to have obtained these results.

## X. Other methods

Other methods were attempted when addressing how to estimate the number of microplastics in the ocean. For various reasons, they did not yield fruitful results, but they are discussed here.

## Kernel Smoothing and Standard Normal Transformation

As mentioned in the Variogram by Year section, there was some evidence to support the idea that different sections of the ocean have different means and variances of microplastic distribution. Intuitively, this also makes sense; sections of ocean that have slow moving water such as ocean gyres may serve as microplastic sinks, as they do for larger pieces of plastic, while section of the ocean with fast-moving currents such as the Atlantic Ocean's Gulf Stream may have fewer pieces of microplastics found in their waters. If this is the case, then it would not make sense to compare the high-microplastic regions directly to the low-microplastic regions. The variance between them would be too large. So, a kernel smoother and standard normal transformation were attempted.

Gaussian smoothers were used for each point after taking a weighted average of the points within a 150-kilometer radius of the original point, a region referred to as the neighborhood. The weight, or kernel, for each of the neighborhood point was calculated with a Gaussian smoother:

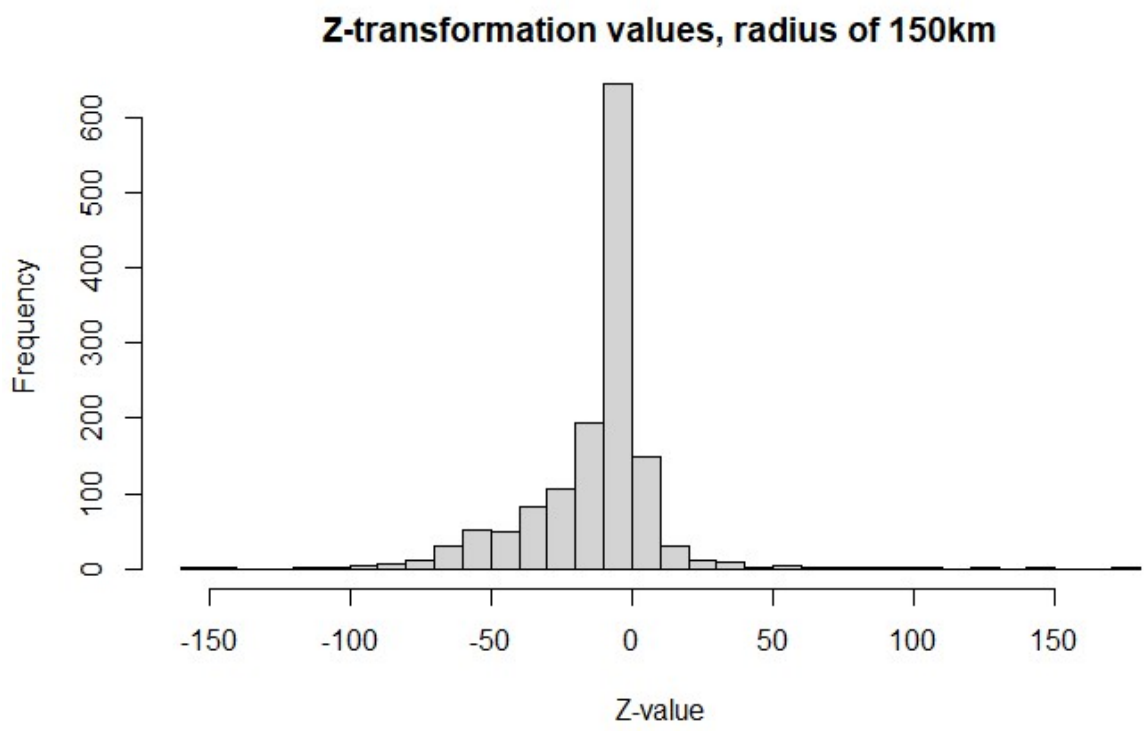
$$K(\tilde{s}, s) = e^{\frac{-1}{2}\left(\frac{\tilde{s}-s}{\delta}\right)^2},$$

where $\tilde{s}$ is the center point, $s$ is a point in the neighborhood, and $\delta$ is the radius of the neighborhood. Weighted averages and variances of the microplastics points for the neighborhood of each point were calculated with this smoother function.

Once each point was transformed to a standard normal distribution using its neighborhood's mean and variance, the histogram of the z-values was plotted. If the transformations worked well, we could expect this histogram to look like a Gaussian distribution centered at 0 with a variance of 1. However, we knew before the graphing of the distribution that it would likely not appear like a $Normal(0,1)$ distribution, owing to the fact that each neighborhood had an average of 11.37 neighbors and that roughly 13% of the data had no neighbors at all. The results are almost comical. With a variance of 546.7, the distribution summary is as follows:

*Z-transformation distribution, radius of 150km*

| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
|---|---|---|---|---|---|
| -150.510 | -19.302 | -4.105 | -11.662 | 0.000 | 177.589 |




Because these results do not approximate a standard normal distribution, this implementation of Gaussian smoothers does not seem to be a good route in order to address varying microplastic distributions globally. Potentially, the Gaussian smoother could be changed to a linear or spherical model. Further, the neighborhood could be widened to include more points in each neighborhood. After all, 150 kilometers was a fairly arbitrary value. However, changing the smoothing function will not be enough to fix the results, and when 300 kilometers was set as a buffer region, the summary statistics for the distribution were similar, and the variance was larger, at 629.8.

*Z-transformation distribution, radius of 300km*

| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
|---|---|---|---|---|---|
| -150.3085 | -26.2504 | -9.9132 | -14.7523 | -0.7587 | 175.6907 |

Because of the failure of the transformed data to conform to a $Normal(0,1)$ distribution, this method of Gaussian smoothing was not pursued further.

## Bayesian Latent Process

Due to the difficulties inherent in the data, the idea of using a Bayesian hierarchical latent process for the data was considered and investigated extensively. This model allows for an underlying

distribution to set the parameter used to generate each data realization, effectively being another way to execute the idea of a normal transformation. In other words, the microplastic distribution Y can be generated when given some underlying process Z. Z is usually defined to be a Gaussian process, and Y|Z can be any other distribution. For our purposes, we would like to use a Poisson or zero-inflated Poisson process. Diggle and Ribeiro demonstrated how to use a Poisson log-linear model with a latent process in their 2007 book Model-based Geostatistics. They applied the process to nuclear radiation values on the island of Rongelap, and a similar process could be used for the microplastics data. This implementation would have used the *geoR* package, which generally requires very specific data formats to work.

Another possibility for implementing the latent process model is with Banerjee, Carlin, and Gelfand's R package *spBayes*. Taking a Bayesian approach, a conditional Poisson model can be implemented, but not a zero-inflated Poisson model. Prior and posterior distributions were written by Prof. Richard Smith, as well as the framework for the iterative process on how to update the parameters and latent process. This was attempted and coded by hand by the author.

Many coding issues were encountered along the way, causing the parameter estimates to not converge. Due to this and a need for finished results, this approach was abandoned. If the currently unknown coding issues can be resolved, however, then this approach would provide a useful comparison with the traditional kriging done above.

## XI. Conclusion

While this model proceeded with many caveats underlying it, the dataset was extremely un-standard for a spatial modeling process. The results may not be entirely reliable, but the fact that estimates were able to be found at all is beneficial for understanding microplastics in the ocean.

Looking to the future of microplastics monitoring, it would be incredibly beneficial if there were a semi-automated process for sampling and counting microplastics. As it is, humans must collect the sample manually, filter it manually, and count all the microplastics on the filter. All three of these steps require time, effort, and logistical coordination. If even one of those steps could be automated, microplastics could be collected and counted easier. It is feasible that a machine learning algorithm could assist in counting the microplastics. Samples of microplastic filters could be photographed, counted, and used as a training set for the algorithm. If the algorithm is able to classify objects as microplastics and then be able to reliably count them, then new samples of microplastics could be processed faster. This would help aid in the data collection process, hopefully allowing more samples to

be taken and increasing the available microplastics data, one of the most important factors for the modeling of their distribution. A similar procedure to this has been done by scientists in Italy, but they used 3D imaging for their machine learning algorithm and were concerned with distinguishing between microplastics and organic diatoms (Bianco, 2019).

With all this being said, the most efficient thing to do to address the issue of ocean microplastics and plastics in general would be to clean up beach plastics first. When plastics are left along the shoreline, it decays and breaks down into microplastics. Higher temperatures from lying on the sand in the sun cause this degradation to happen more quickly than in the ocean, as well as higher oxygen concentrations (Andrady, 2011), making beach cleanups a critical component for preventing the creation of new microplastics. Shorelines are much easier to access than the open ocean, and clean-ups can be done by volunteers with no prior experience, another reason for focusing on beach cleanups. Tomoya and Hirofumi showed in their 2015 paper that there are times when a beach cleanup's effect will be greatest, chiefly when the amount of plastics on the beach reaches a local maximum. Hence, the timing of beach cleanups can be optimized to enhance beach cleanup efforts, thus "decreasing the total mass of toxic metals that could leach into the beach from marine plastics and prevent the fragmentation of marine plastics" (Tomoya and Hirofumi, 2015).

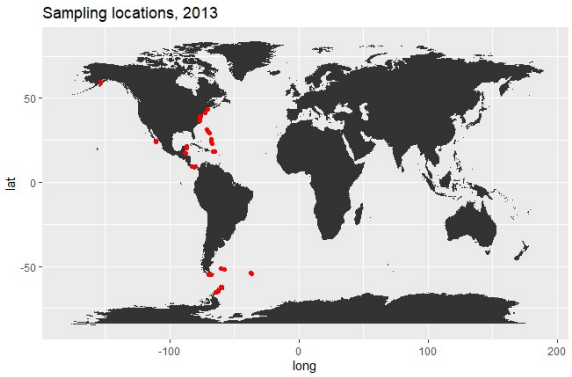
Microplastics are an important source of pollution that should be spatially modeling in order to identify regions of accumulation and to best implement cleanup solutions. Because of the high number of samples with no microplastics, a zero-inflated Poisson distribution may prove to be better at modeling the microplastic distribution than the methods used here. Further, the parameters used for the preliminary variogram and kriging model should be reevaluated in order to find the best fit. This paper does show that there is evidence for spatial correlation in microplastic distribution and that kriging methods could be used to model their distribution.

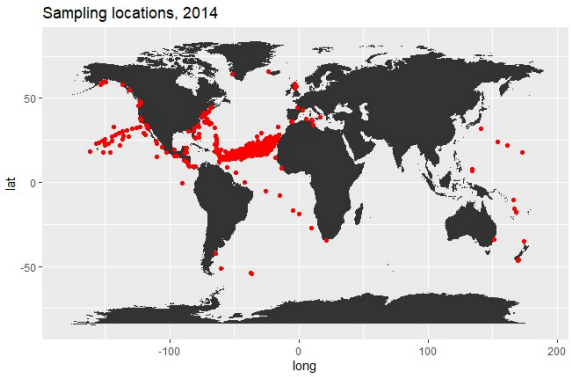
## Appendix A: Code for the Variogram Function

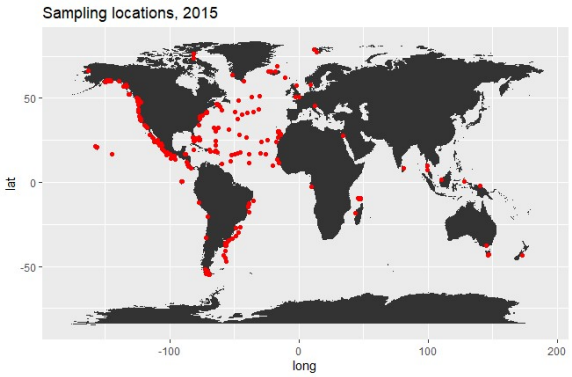
The following R code was used to create the variograms seen in this report.

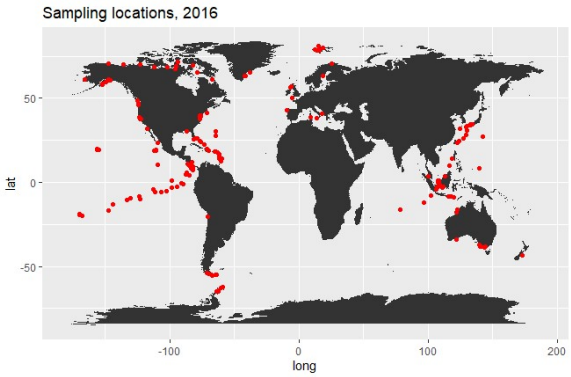
```
library(geosphere)

variogram = function(coords_in, mp_in, low, high, interval, delta,
main){
  output1 = c()

  coords = coords_in
  mp = as.numeric(c(mp_in))

  for (h in seq(low,high,interval)){
    if((h/10)%%5 == 0){
      print(h)
    }
    sum1=0
    sum2=0
    for (i in (1:nrow(coords))){
      distv = as.numeric(distm(c(coords$long[i], coords$lat[i]),
data.frame(coords$long, coords$lat), distHaversine))
      distv=distv/1000
      distvplus=data.frame(distv, mp)
      distvsub = subset(distvplus, distv>h-delta & distv<h+delta)
      sum1=sum1+nrow(distvsub)
      if (nrow(distvsub) != 0){
        sum2 = sum2+sum((distvsub$mp-coords$pieces.per.L[i])^2)
      }
    }
    value1 = sum2/sum1
    output1 = c(output1, value1)
  }
  plot(seq(low, high, interval), output1, main = main)
}
```

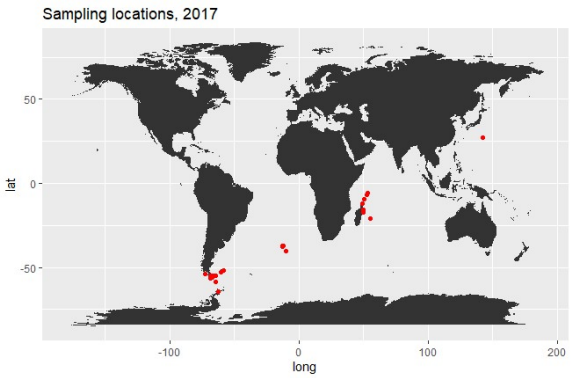
# Appendix B: Maps of sampling locations by year



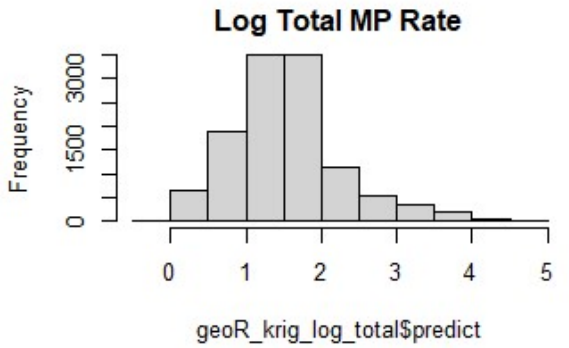



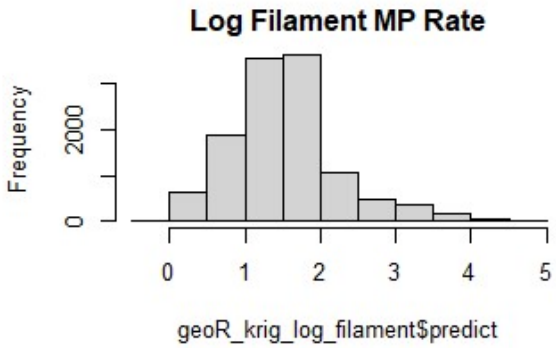



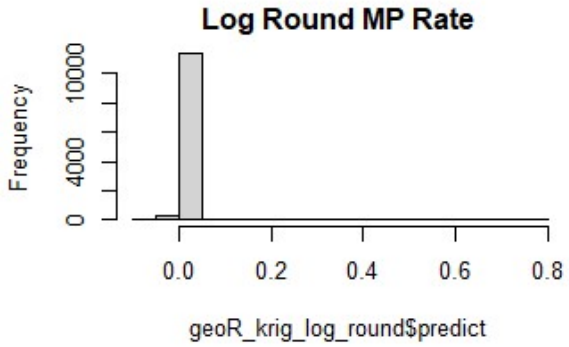



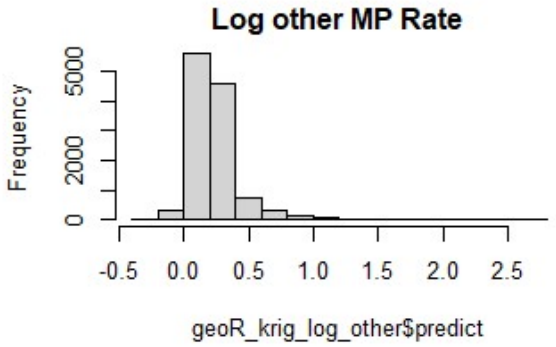



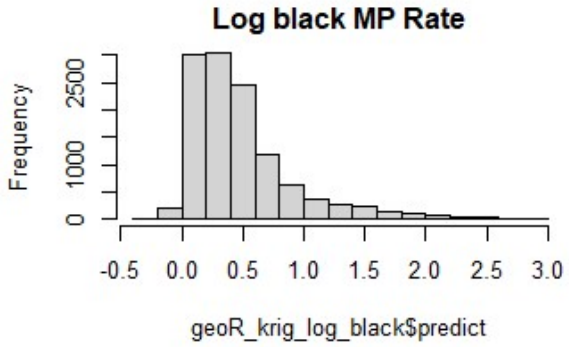

# Appendix C: Results by variable type

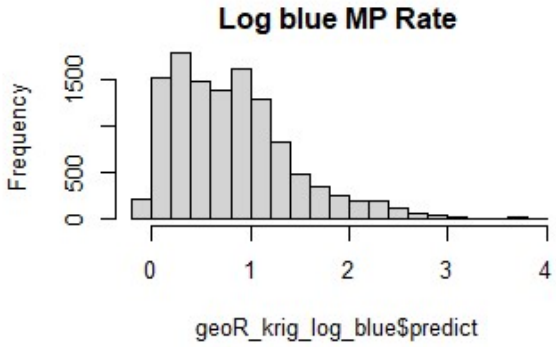

Log Total MP Rate
Frequency
0
1500
3000
0
1
2
3
4
5
geoR_krig_log_total$predict

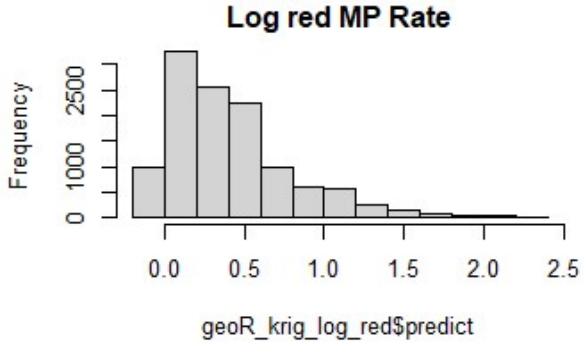

Log Filament MP Rate
Frequency
0
2000
0
1
2
3
4
5
geoR_krig_log_filament$predict

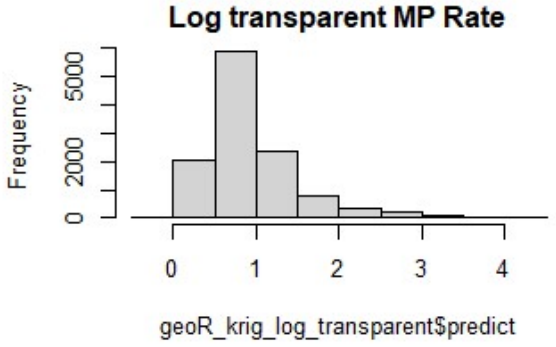

Log Round MP Rate
Frequency
0
4000
10000
0.0
0.2
0.4
0.6
0.8
geoR_krig_log_round$predict

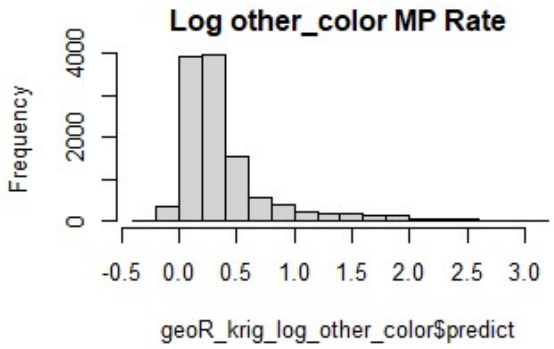

Log other MP Rate
Frequency
0
2000
5000
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
geoR_krig_log_other$predict

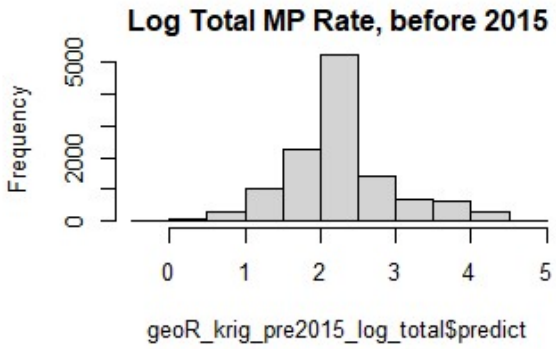

Log black MP Rate
Frequency
0
1000
2500
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
geoR_krig_log_black$predict

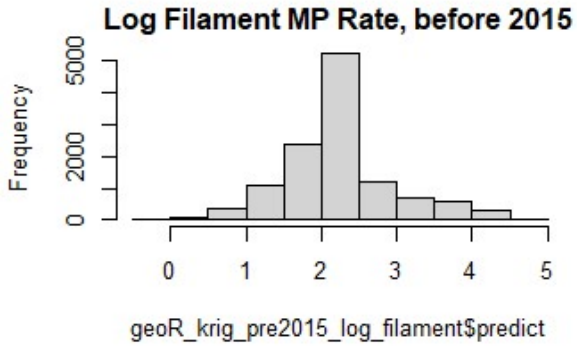

Log blue MP Rate
Frequency
0
500
1500
0
1
2
3
4
geoR_krig_log_blue$predict

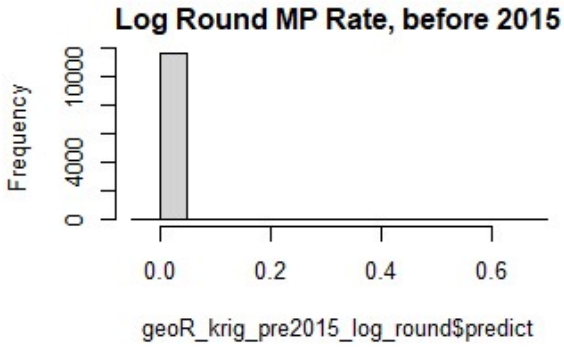

Log red MP Rate
Frequency
0
1000
2500
0.0
0.5
1.0
1.5
2.0
2.5
geoR_krig_log_red$predict

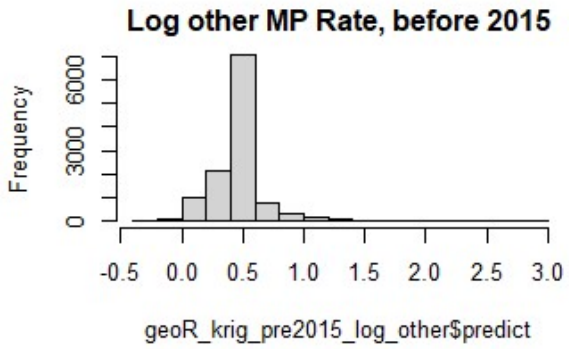

Log transparent MP Rate
Frequency
0
2000
5000
0
1
2
3
4
geoR_krig_log_transparent$predict

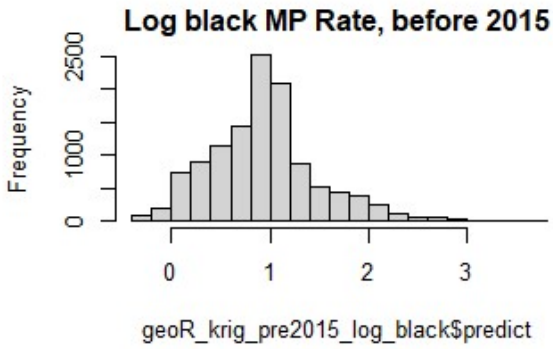

Log other_color MP Rate
Frequency
0
2000
4000
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
geoR_krig_log_other_color$predict

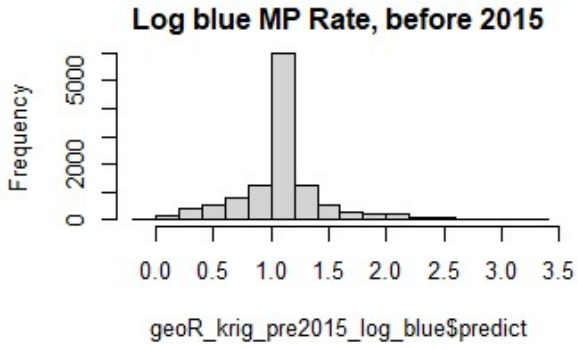

Log Total MP Rate, before 2015
Frequency
0
2000
5000
0
1
2
3
4
5
geoR_krig_pre2015_log_total$predict

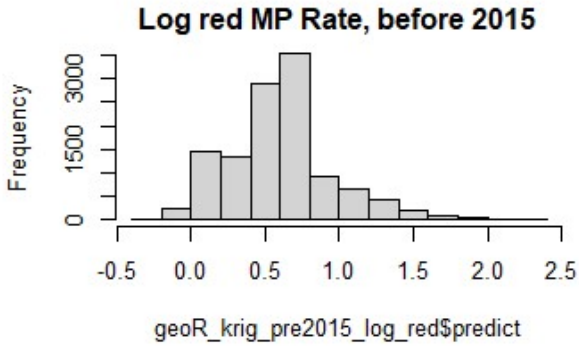

Log Filament MP Rate, before 2015
Frequency
0
2000
5000
0
1
2
3
4
5
geoR_krig_pre2015_log_filament$predict

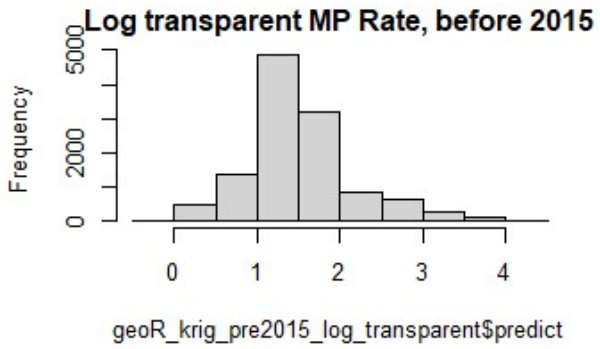

Log Round MP Rate, before 2015
Frequency
0
4000
10000
0.0
0.2
0.4
0.6
geoR_krig_pre2015_log_round$predict

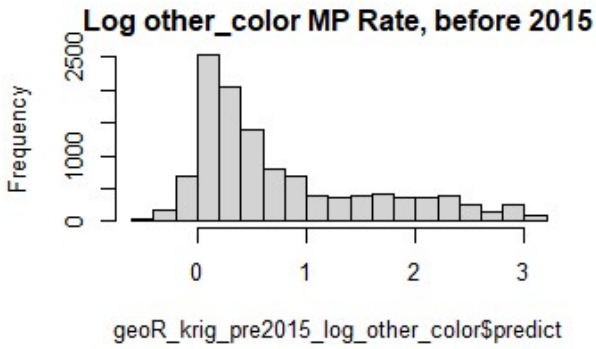

Log other MP Rate, before 2015
Frequency
0
3000
6000
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
geoR_krig_pre2015_log_other$predict

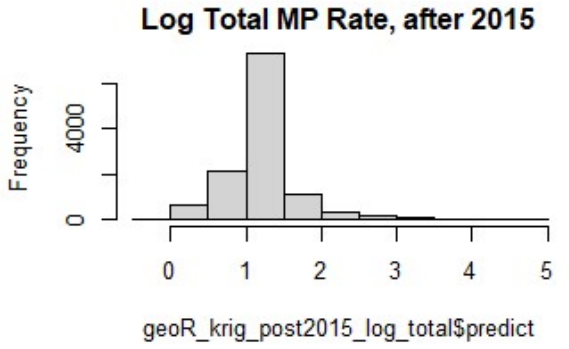

Log black MP Rate, before 2015
Frequency
0
1000
2500
0
1
2
3
geoR_krig_pre2015_log_black$predict

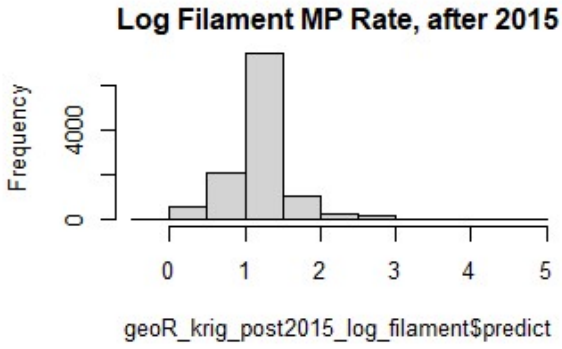

Log blue MP Rate, before 2015
Frequency
0
2000
5000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
geoR_krig_pre2015_log_blue$predict

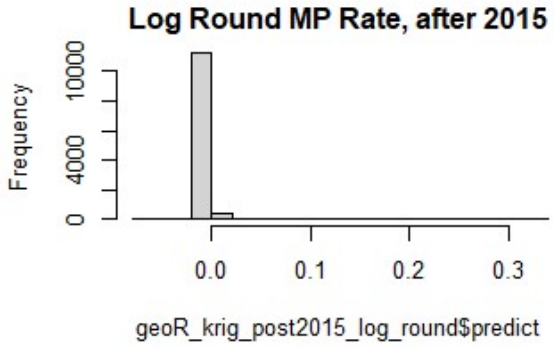

Log red MP Rate, before 2015
Frequency
0
1500
3000
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
geoR_krig_pre2015_log_red$predict

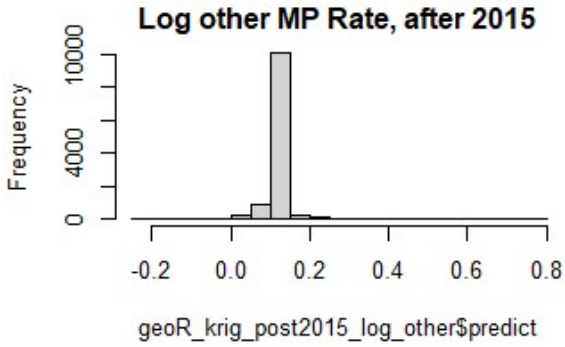

Log transparent MP Rate, before 2015
Frequency
0
2000
5000
0
1
2
3
4
geoR_krig_pre2015_log_transparent$predict

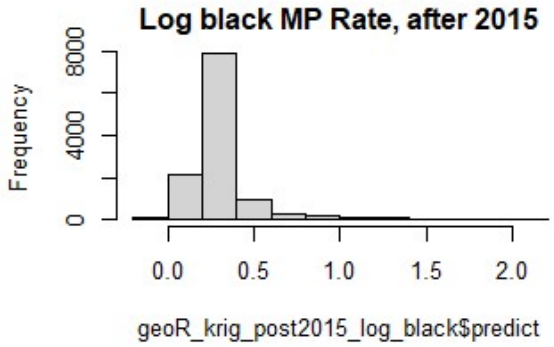

Log other_color MP Rate, before 2015
Frequency
0
1000
2500
0
1
2
3
geoR_krig_pre2015_log_other_color$predict

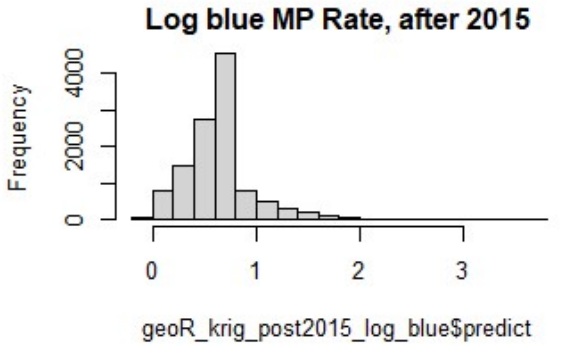

Log Total MP Rate, after 2015
Frequency
0
4000
0
1
2
3
4
5
geoR_krig_post2015_log_total$predict

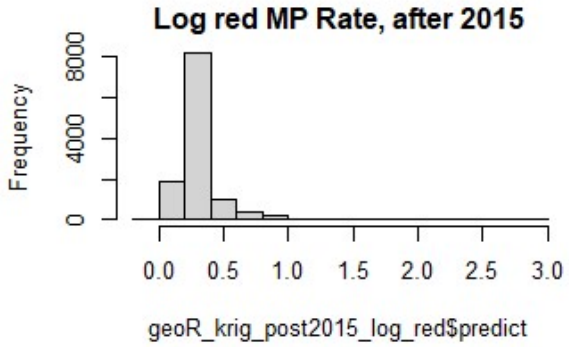

Log Filament MP Rate, after 2015
Frequency
0
4000
0
1
2
3
4
5
geoR_krig_post2015_log_filament$predict

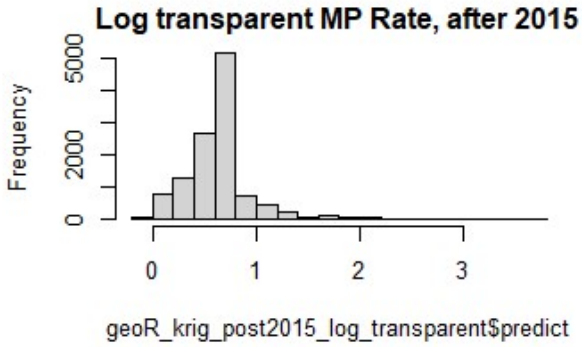

Log Round MP Rate, after 2015
Frequency
0
4000
10000
0.0
0.1
0.2
0.3
geoR_krig_post2015_log_round$predict

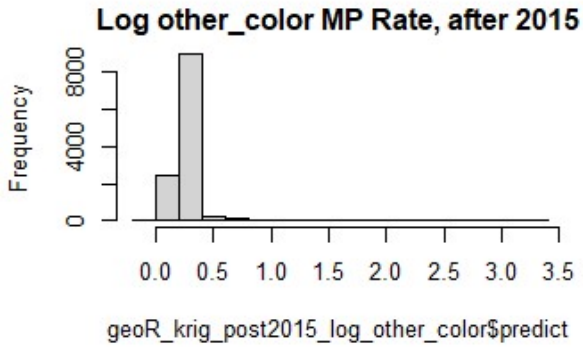

Log other MP Rate, after 2015
Frequency
0
4000
10000
-0.2
0.0
0.2
0.4
0.6
0.8
geoR_krig_post2015_log_other$predict


Log black MP Rate, after 2015
Frequency
0
4000
8000
0.0
0.5
1.0
1.5
2.0
geoR_krig_post2015_log_black$predict


Log blue MP Rate, after 2015
Frequency
0
2000
4000
0
1
2
3
geoR_krig_post2015_log_blue$predict


Log red MP Rate, after 2015
Frequency
0
4000
8000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
geoR_krig_post2015_log_red$predict


Log transparent MP Rate, after 2015
Frequency
0
2000
5000
0
1
2
3
geoR_krig_post2015_log_transparent$predict


Log other_color MP Rate, after 2015
Frequency
0
4000
8000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
geoR_krig_post2015_log_other_color$predict

# Appendix D: References

Andrady, A. L. *Microplastics in the marine environment.* Marine Pollution Bulletin, Volume 62, Issue 8, pages 1596-1605. 2011.

Asmarian, Naeimehossadat et al. "Bayesian Spatial Joint Model for Disease Mapping of Zero-Inflated Data with R-INLA: A Simulation Study and an Application to Male Breast Cancer in Iran." *International journal of environmental research and public health* vol. 16,22 4460. 13 Nov. 2019.

Banerjee, S., Carlin, B.P., Gelfand, A.E. Hierarchical Modeling and Analysis for Spatial Data, Second Edition. CRC Press, Taylor and Francis Group, Boca Raton, FL. 2015.

Barrows, A.P., Neumann, C.A., Berger, M.L., Shaw, S.D. *Grab vs. neuston tow net: a microplastic sampling performance comparison and possible advances in the field.* Analytical Methods, Volume 9, Issue 9, pages 1446-1453. 2017.

Barrows, A.P.W. et al. *Marine environment microfiber contamination: Global patterns and the diversity of microparticle origins.* Environmental Pollution, Volume 237, pages 275-284. 2018.

Bianco, V., Pasquale, M., et al. *Microplastic Identification via Holographic Imaging and Machine Learning.* Advanced Intelligent Systems, Volume 2, Issue 2. 10 December 2019.

Diggle, Peter, and Paulo J. Ribeiro. *Model-based Geostatistics*. New York, NY: Springer, 2007. Internet resource.

Eriksen, M., Lebreton, L., et al. *Plastic Pollution in the World's Oceans: More than 5 trillion Plastic Pieces Weighing over 250,000 Tons Afloat at Sea*. Plos One. 2014.

Hardesty, B.D., Harari, J., Isobe, A., Lebreton, L., et al. *using Numerical Model Simulations to Improve the Understanding of Micro-plastic Distribution and Pathways in the Marine Environment.* Frontiers in Marine Science, Volume 4, page 30. 2017.

Iñiguez, M.E., Conesa, J.A., Fullana, A. *Microplastics in Spanish Table Salt*. Scientific Reports, Volume 7, page 8620. 2017.

Isobe, A., Uchiyama-Matsumoto, K., Uchida, K., Tokai, T. *Microplastics in the Southern ocean.* Marine Pollution Bulletin, Volume 114, pages 623-626. 2017.

Kennedy, Dana. “Study Finds Massive Pollution from Microfibers in California Oceans, Waterways.” *New York Post*, New York Post, 17 Oct. 2020, nypost.com/2020/10/17/study-finds-huge-pollution-from-microfibers-in-californias-oceans/.

Leonard, George, et al. “Plastics in the Ocean.” *Ocean Conservancy*, 30 Oct. 2020, oceanconservancy.org/trash-free-seas/plastics-in-the-ocean/.

“Machine Learning Scopes out Previously ‘Invisible’ Microplastics.” *Advanced Science News*, 4 Feb. 2020, www.advancedsciencenews.com/machine-learning-scopes-out-previously-invisible-microplastics/.

Nychka, Douglas. “Package ‘fields” Reference Manual. October 2020. https://cran.r-project.org/web/packages/fields/fields.pdf

Nychka, Douglas. “SpatialProcess: Estimates a Spatial Process Model. in Fields: Tools for Spatial Data.” *R Package Documentation*, 23 Oct. 2020, rdrr.io/cran/fields/man/spatialProcess.html.

Tomoya Kataoka and Hirofumi Hinata. *Evaluation of beach cleanup effects using linear system analysis.* Marine Pollution Bulletin, Volume 91, Issue 1, pages 73-81. 2015.

UNCLOS, 1982. United nations convention on the law of the sea. *Division for Ocean Affairs and the Law of the Sea*.

Wieczorek, A., Morrison, L., Croot, P., Allcock, L., et al. *Frequency of Microplastics in Mesopelagic Fishes from the Northwest Atlantic*. Frontiers in Marine Science, Volume 5, page 39. 2018.

Wilcox, C., Van Sebille, E., Hardesty, B.D., *Threat of plastic pollution to seabirds is global, pervasive, and increasing.* Proceedings of the National Academy of Sciences Early Edition, 11899-11904. 2015.