

Distribution Fitting of a Regular Point Process

Leslie Morgan^{†‡}, Andrew Martinez[‡], Leann Myers[‡],
Brian Bourgeois[†]

I. Introduction

The Naval Research Laboratory at Stennis Space Center, MS is developing a terrain-based navigation system that uses multibeam bathymetry to estimate the position of an autonomous underwater vessel (AUV). A maximum likelihood approach is used to find the most likely position of the vessel based upon the vehicle's last estimated position, its current measured ocean depth and a bathymetry map of the area [1]. The bathymetry point from the map that most closely matches the vessel measured ocean depth is used as the estimate of the vessel's position.

A parameterized model is needed to quantify the lower bound on the estimated position error for the terrain-based navigation system being developed. The current research is concerned with the characterization of the point pattern produced by multibeam sonar systems and with the development of a parameterized model for the point-to-event distance distribution. The parameterized model will provide confidence intervals for the expected distance to the nearest bathymetry point from any arbitrary point and give an estimate of the average positioning error that would be observed if the system always picked the nearest bathymetry point from the vessel's true location.

The proposed model was developed by first testing the null hypothesis of complete spatial randomness on a sample bathymetric data set. The null hypothesis of complete spatial randomness, the tests used to evaluate this null hypothesis, and the results of these tests are discussed in the next section. The third section describes the process used to select an appropriate parameterized point-to-event model. A two-parameter Weibull distribution was found to provide a reasonably good fit to the observed data. The fourth section investigates the generalizability of the Weibull model. The last section discusses future work and summarizes the findings to date.

II. Null Hypothesis Testing

A. Data Description and the Null Hypothesis

The dataset used to develop the parameterized model was obtained from Pensacola Bay, Florida. The data are not gridded and contain 29,963 (x, y, z)

coordinate points in meters. The x-values range from 400m to 1100m and the y-values from 300m to 775m. For the current analyses, only interior regions of the dataset were used to avoid edge effects. This is a valid restriction in that terrain-based navigation systems should not be used on the edge of available bathymetry data.

The (x, y) coordinate points for the multibeam bathymetry data produce a two-dimensional point pattern. The null hypothesis of complete spatial randomness (CSR) assumes that a homogenous planar Poisson point process produced this point pattern. The null hypothesis of complete spatial randomness serves as a dividing line for the alternatives of regularly spaced or clustered patterns. Because one of the goals for the production of multibeam bathymetry data is to achieve a regular sampling of points, one would expect to be able to reject the null in favor of a regular alternative. Under the null hypothesis, the number of points in any set A will have a Poisson distribution with mean λ . λ is a constant which represents the intensity of the process. An unbiased estimator of λ is the number of points in the study region divided by the area of the study region [2].

There is no single definitive test of the null hypothesis of CSR. There are several recommended methods, but formal comparative studies are few. The power of the different available tests varies according to the type of pattern under observation [3]. Diggle [4] suggests that several different tests should be used to provide both complementary evidence for a conclusion and to reveal various attributes of the pattern through different analyses. Three analyses were performed on the multibeam data. These are the refined nearest neighbor analysis, a second-order analysis via the K-function, and a point-to-event analysis. Each of these analyses and their results are discussed in the remainder of this section.

B. Nearest Neighbor Analysis

The refined nearest neighbor analysis involves comparing the cumulative distribution function of all the nearest neighbor distances within the study area with the expected cdf under CSR. Under CSR, the expected cdf is $G(r) = 1 - \exp(-\lambda\pi r^2)$, $r \geq 0$, where r is the distance to the nearest neighbor. Lambda (λ) is once again defined as the intensity of the process and is estimated by N/A where N is the number of points in the study area and A is the size of the study area. The region $450m \leq X \leq 550m$ and $350m \leq Y \leq 450m$ was used for this analysis. For this study area, $\hat{\lambda} = 0.1114$. The observed and expected cdf's are shown in Figure 1.

[†] Naval Research Laboratory, Stennis Space Center, Mississippi

[‡] Tulane University, New Orleans, Louisiana

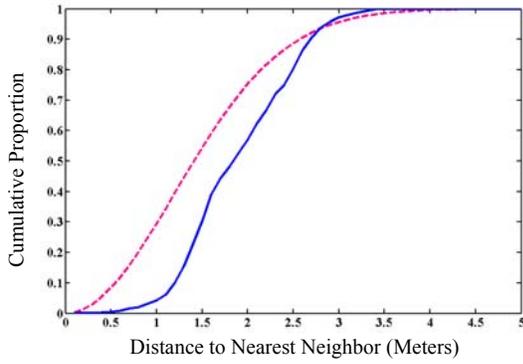


Fig. 1. Observed and expected nearest neighbor cumulative distribution functions (cdf). The lower solid curve shows the cdf for the multibeam bathymetry data. The upper dashed curve shows the expected cdf under CSR.

The lower cdf is that of the observed nearest neighbor distances for the multibeam bathymetry data in the study region. The upper cdf is the expected cdf under the null hypothesis generated by $G(r)$ with intensity $\hat{\lambda}$. That the observed cdf is lower than the expected cdf is evidence of a regular process [3]. The multibeam cdf lags until a radial distance of about one meter between events is reached. After one meter between events is reached, the multibeam cdf begins to grow exponentially. This suggests an inhibition distance of approximately one-meter between points.

Figure 1 shows an empirical difference between the cdf for the multibeam data and the cdf expected under the null hypothesis. A formal test is needed to determine whether the observed and expected cdf's differ significantly. Second-order analysis provides this formal test.

C. Second-Order Analysis

Because the bathymetry data are an exhaustive map of all the bathymetry points within the study area, second-order analysis could be performed. Second-order analysis is the study of inter-event distances, where the events are mapped points. Second-order analysis estimates the K-function. The K-function is closely related to the second-order intensity of a stationary isotropic process, and for this reason, is often called the reduced second moment measure [5]. The advantages of this type of analysis are that it reveals spatial information at all scales of pattern and the exact locations of all events are used in the estimation.

A detailed explanation of the K-function can be found in Cressie [5]. For a Poisson process, $E(K(r)) = \pi r^2$, where r is distance. For a regular process, $\hat{K}(r)$ will be less than πr^2 , and for a clustering process, it will

be greater than πr^2 . For simplification, the plot of $\hat{K}(r)$ for a Poisson process can be linearized by the function $\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}}$, making $E(\hat{L}(r)) = r$. This linearization also has the effect of stabilizing the variances [2].

$\hat{K}(r)$ for the multibeam bathymetry data was estimated in the study region $450m \leq X \leq 550m$ and $350m \leq Y \leq 450m$. Figure 2. shows $\hat{L}(r) - r$ versus upper and lower envelopes from 100 simulations of a Poisson process in the study area. $\hat{L}(r) - r$ is plotted every 0.25 meters for 0.25 meters $\leq r \leq 30$ meters.

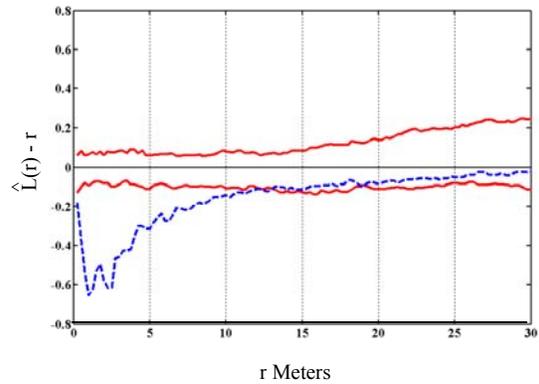


Fig. 2. 2nd Order analysis results. The dashed curve is $\hat{L}(r) - r$ for the bathymetry data. The solid curves show the upper and lower envelopes of $\hat{L}(r) - r$ for 100 simulated Poisson processes. The expected value of $\hat{L}(r) - r$ for a Poisson process is zero.

Examination of Figure 2 shows that the bathymetry data differ significantly from a Poisson process on a scale of r less than approximately 12 meters. This graph also reveals an inhibition distance of approximately one-meter, evidenced by the point at which the curve's slope first becomes positive. This is consistent with the observation of an inhibition distance of about one-meter noted in the refined nearest neighbor analysis. This inhibition distance means that it is unlikely for two bathymetry points to be closer than one-meter to each other.

D. Point-To-Event Distance Analysis

The point-to-event distance analysis is related to the refined nearest neighbor analysis. This analysis measures the distance from each of m sample points to the closest of the n events in the study area. The m sample points are placed randomly in the study area based on a jointly uniform distribution. From these distances, the cumulative distribution function for the

point to nearest event distances, $F(r)$, is estimated. Under CSR, $F(r) = 1 - \exp(-\lambda\pi r^2)$, where $r \geq 0$ and λ is the number of events divided by the study area.

The cdf's for the point to nearest event distances were estimated for five separate 50 by 50-meter regions of the multibeam bathymetry data. These cdf's were estimated by placing 2500 points from a jointly uniform distribution within each of the five regions. These regions, the number of multibeam bathymetry points per region, n , and the intensity of the process within each region, $\hat{\lambda}$, are shown in Table 1.

	Region 1	Region 2	Region 3	Region 4	Region 5
n	283	275	281	280	275
$\hat{\lambda}$	0.1132	0.1100	0.1124	0.1120	0.1100

Table 1. Point-to-event distance study region parameters

The empirical cdf's for the point to nearest event distances for the five separate regions are shown by the solid curves in Figure 3. The expected cdf under CSR with λ estimated by the average $\hat{\lambda}$'s from the five regions is shown by the dashed curve.

The observed cdf's are above the cdf expected under CSR. This is once again evidence of regular spacing of the data [3]. The five cdf's were not found to differ significantly, (k-sample Kolmogorov-Smirnov test, $p > 0.10$).

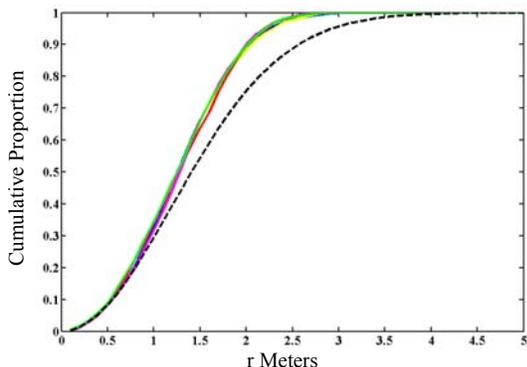


Fig. 3. The empirical point-to-event distance cdf's for the five study areas are shown by solid lines. The dashed curve is the expected cdf under CSR.

E. Results of Null Hypothesis Testing

Based on the results of the above analyses, there was sufficient evidence to reject the null hypothesis of complete spatial randomness in favor of a regularly spaced alternative hypothesis. The multibeam bathymetry data exhibit greater regularity than that

expected from a homogenous planar Poisson process. The nearest neighbor analysis and the second-order analysis show an inhibition distance of approximately one-meter.

The results of the previous analyses show that the point-to-event distributions of the five study regions do not differ at the 0.10 level of significance. This implies that one parametric model for the point-to-event distance could be used to fit the entire data set. The next section is concerned with the development of an appropriate model.

III. Parameterized Model Selection

A parameterized model for the distribution of the point-to-event distances within the multibeam bathymetry data is needed to estimate the lower limit of the positioning error for the terrain-based navigation system being developed. This parameterized distribution should have a lower limit of zero since there will be no negative distances from an arbitrary point to the closest bathymetric point. Transforming the data from linear distances to circular areas is logical because as one moves outward from an arbitrary point in search of the nearest event, one is moving outward along a radius that encloses a circular region around the point.

When the data are completely spatially random, the point-to-event distance distribution is exponential [5]. An exponential distribution is characterized by a constant hazard rate. That is, from any arbitrary point, the probability of encountering an event is the same for all radial distances from that point. The multibeam data exhibits regularity and an inhibition distance, so a distribution with an increasing hazard rate is expected. From an arbitrary point in the multibeam data, the chance of encountering an event may be low at first if the arbitrary point is within the inhibition distance between two points. The observed regular spacing of the events ensures that as one moves out further and further along the radial line, the chance of encountering a multibeam bathymetry point increases.

The five empirical cdf's for the point-to-event distances that were estimated in Section 2 and shown in Figure 3 were used to determine an appropriate parametric distribution. The first step in determining an appropriate model is to select a group of potential models. Given that the distribution should have a lower bound of zero, and an increasing hazard rate, six standard survival distributions were selected as potential candidates: the generalized gamma, the exponential, the Weibull, the standard gamma, the log-normal, and the log-logistic distributions. Typically, survival distributions are used to estimate time-to-event. These distributions were appropriate because the

radial distance-to-event can be thought of as analogous to time-to-event. The procedure Proc Lifereg in SAS (Statistical Analysis System) was used to determine the log-likelihood for each of the distributional models in each of the five study regions. The log-likelihoods for the generalized gamma, Weibull, gamma and exponential distributions are shown in Table 2. The log-likelihoods for the log-normal and the log-logistic distributions are not shown due to their lack of fit. Lower magnitudes correspond to better fits. In all areas assessed, the generalized gamma provided the best fit followed by the Weibull distribution.

	Region 1	Region 2	Region 3	Region 4	Region 5
Generalized Gamma	-3511	-3580	-3485	-3412	-3549
Weibull	-3516	-3592	-3513	-3433	-3576
Gamma	-3528	-3604	-3534	-3458	-3594
Exponential	-3578	-3638	-3583	-3525	-3630

Table 2. Estimated log-likelihoods for each distribution type.

The generalized gamma is a three-parameter distribution involving the gamma function and the incomplete gamma function. The exponential, Weibull, standard gamma, and log-normal models are all special cases of the generalized gamma distribution. The third parameter of the generalized gamma allows its hazard function to take on a wide variety of shapes. The generalized gamma distribution will fit unless the hazard function has more than one peak [6]. Hazard function plots of the multibeam data did not reveal a distribution with more than one peak. If one of the simpler models can be shown to fit, the generalized gamma is not used for three main reasons. The pdf is complicated, and the parameters are difficult to interpret. The computer time to estimate the generalized gamma is significantly longer than for the simpler models. The generalized gamma has a reputation for convergence problems [6]. For these reasons, the Weibull distribution was selected as the potential model.

The Weibull model is a slight modification of the exponential model, with the important consequence that the hazard rate is no longer constant. The Weibull cdf incorporating the transformation to radial areas is given by $G(r) = 1 - \exp(-\lambda(\pi r^2)^\gamma)$, $r \geq 0$, $\lambda > 0$, and $\gamma > 0$. λ is a scale parameter and γ is a shape parameter. When $\gamma = 1$, the Weibull cdf reduces to the exponential. When $1 < \gamma < 2$, which is the case for the multibeam point-to-event distributions, the hazard rate is increasing at a decreasing rate [6]. For the Weibull distribution given above, the expected value of r is

$$\text{given by } E(r) = \frac{1}{\sqrt{\pi}} \lambda^{-1/2\gamma} \Gamma\left(1 + \frac{1}{2\gamma}\right) \text{ and the variance is}$$

$$\text{given by } \text{Var}(r) = \frac{\lambda^{-1/\gamma}}{\pi} \left[\Gamma\left(1 + \frac{1}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{2\gamma}\right) \right].$$

Weibull probability plots are shown in Figure 4. The plots of the empirical distributions from the five study regions are shown, as well as the plot of a simulated Weibull distribution.

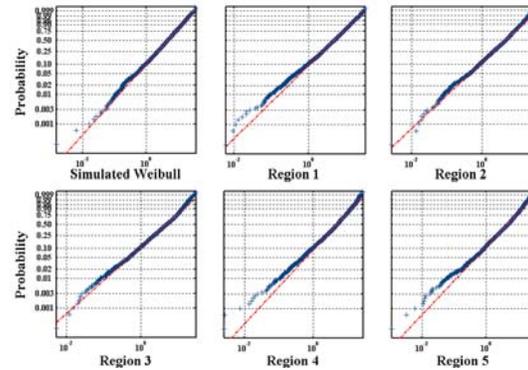


Fig. 4. Weibull probability plots for a simulated Weibull distribution and the empirical distributions for the 5 study areas.

The method of maximum likelihood was used to estimate the parameters for a Weibull distribution for each of the five empirical cdf's. These estimated parameters and their 95% confidence intervals are shown in Table 3.

	λ	95% CI	γ	95% CI
Region 1	0.1052	(0.1010, 0.1095)	1.2016	(1.1647, 1.2385)
Region 2	0.1118	(0.1072, 0.1165)	1.1715	(1.1337, 1.2093)
Region 3	0.1023	(0.0979, 0.1067)	1.2192	(1.1795, 1.2588)
Region 4	0.0950	(0.0910, 0.0990)	1.2575	(1.2184, 1.2965)
Region 5	0.1112	(0.1065, 0.1159)	1.1901	(1.1525, 1.2276)

Table 3. Estimates of Weibull distribution parameters for the five study regions.

To determine goodness-of-fit of the Weibull model, two-sample Kolmogorov-Smirnov tests were performed. One thousand Weibull distributions of 2500 observations using the estimated parameters were simulated for each of the five empirical cdf's. The procedure Proc NPAR1Way in SAS was used to obtain the maximum distance, D , between the two cdf's. The minimum D , the maximum D , and the mean of D 's, for each of the five regions are shown in Table 4. This table also contains the results of one simulated Weibull distribution compared to 1000 other simulated Weibull

distributions with the same parameters. The average maximum distance between the cdf's suggested that the Weibull distribution would be an adequate approximation to the multibeam point-to-event distribution.

	Weibull	Region 1	Region 2
Maximum D	0.0464	0.0512	0.0600
Minimum D	0.0100	0.0108	0.0128
Average D	0.0242	0.0227	0.0274
	Region 3	Region 4	Region 5
Maximum D	0.0696	0.0676	0.0772
Minimum D	0.0212	0.0168	0.0136
Average D	0.0425	0.0339	0.0363

Table 4. Range of D values and average D value for the empirical cdf's versus the modeled cdf's, and for the simulated Weibull distribution.

IV. Generalizability of the Weibull Model

To be a reliable estimate of positioning error for the terrain-based navigation system, the Weibull model needs to be generalizable to other multibeam data sets. To determine under what conditions a Weibull model would provide adequate approximation, a point pattern process that replicates the multibeam pattern was needed. The goal of the multibeam process is to produce a regular grid of points, but due to the nature of the process, a certain amount of noise is introduced. A regular pattern of points with random noise introduced to each point was generated to replicate the multibeam pattern.

To determine if a regular process with random noise adequately describes the multibeam process, regular patterns with noise were created by simulating 50 by 50 meter grids of events. The events were placed three meters apart to replicate the multibeam density, approximately 0.11 events per square meter. A random amount of noise from a jointly uniform distribution ($-h \leq x \leq h$, $-h \leq y \leq h$, $h = \text{noise in meters}$) was then introduced to each event. Regular patterns with noise equal to 1.5 meters provided the closest replication to the Pensacola Bay multibeam data. The K-function was used to determine if the simulated process fit the multibeam process.

Figure 5 shows upper and lower envelopes from 100 simulations of regular patterns with noise equal to 1.5 meters versus the linearized K-function for three 50 by 50 meter regions of the multibeam data. The fit is not perfect but close. The multibeam data appears to have a slightly greater inhibition distance than the pattern with noise.

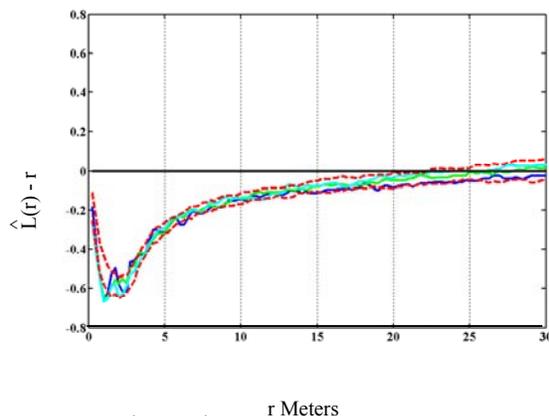


Fig. 5. Plot of $L(r) \equiv (K(r)/\pi)^{1/2} - r$, versus upper and lower envelopes from 100 simulations of regular processes with noise = 1.5 meters. The upper and lower envelopes are represented by the dashed lines.

A regular pattern with noise equal to 1.5 meters was determined to adequately describe the multibeam pattern. The next step was to both confirm that the Weibull model fit the point-to-event distribution of the simulated pattern and to determine if the Weibull model was generalizable to other noise levels. Regular patterns with noise were produced with noise increasing from 0.0 meters to 6.0 meters in increments of 0.1 meters. Ten simulations were done for each increment in noise. Empirical point-to-event distributions were obtained by randomly placing 2500 points within the 50 by 50 meter grids and by measuring the distance from each point to the nearest event.

Proc Lifereg in SAS was used to obtain log-likelihoods for the generalized gamma distribution, the Weibull distribution, and the exponential distribution for each of the point-to-event distributions. The ten log-likelihoods for each distribution type were averaged at each level of noise. The average Weibull log-likelihood was subtracted from the average generalized gamma log-likelihood to produce a chi-square statistic on one degree of freedom. Likewise, the average exponential log-likelihood was subtracted from the average Weibull log-likelihood.

The results of the log-likelihood differences are shown in Figure 6. Although the generalized gamma would fit better, Kolmogorov-Smirnov tests revealed the Weibull distribution to provide a reasonably good fit for noise levels greater than or equal to 1.5 meters, whereas the exponential did not begin to fit until noise levels reached 2.7 meters.

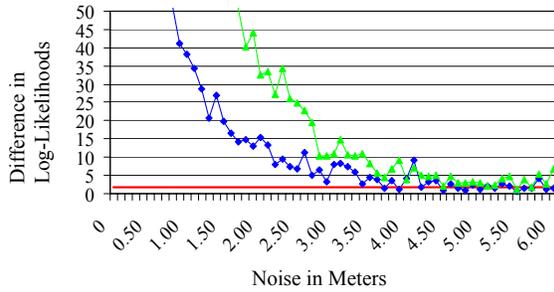


Fig. 6. Average difference in log-likelihoods from ten point-to-event simulations at each noise increment. The uppermost curve is the Weibull minus the exponential. The lower curve is the generalized gamma minus the Weibull. The solid line is $y = 1.64$ which represents a p-value of 0.20 for a chi-square statistic on one degree of freedom.

V. Conclusions and Future Work

For the multibeam data set studied, the null hypothesis of CSR was rejected in favor of a regular alternative. The average density of the multibeam points is approximately 1 point per 9 square meters. There is an inhibition distance between events of about 1.1 meters. The point-to-event data were adequately modeled by a Weibull distribution. The parameters for this model can be easily and reliably estimated.

A regular point pattern with random noise of 1.5 meters introduced to each point was found to replicate the multibeam pattern under study. The Weibull model was found to adequately fit the point-to-event distribution for regular patterns with noise greater than or equal to 1.5 meters. That is, when the noise was greater than or equal to half the average distance between points, the Weibull model fit. The Weibull model fit significantly better than an exponential model until noise levels reached approximately 2.7 meters.

Future work will involve using the Weibull model to estimate the lower bound on positioning error for the terrain-based navigation system. The analysis of other multibeam data sets will be conducted to determine if the same regularity of pattern exists and to determine if the Weibull model can be generalized to these data sets.

Acknowledgments

This work was funded by the Office of Naval Research through the Naval Research Laboratory under Program Element 62435N. The mention of commercial products or the use of company names does not in any way imply endorsement by the U.S. Navy. Approved for public release; distribution is unlimited. NRL contribution number NRL/PP/7440-01-1006.

References

- [1] Beckman, Richard, Martinez, Andrew, and Bourgeois, Brian (2001), "LOST2: A Terrain Based Underwater Positioning System; Results from Sea Trials," *Proceedings of the 12th Intl. Symposium on Unmanned Untethered Submersible Technology*, 27-29AUG01, Durham, NH.
- [2] Ripley, Brian D. (1981), *Spatial Statistics*, New York: John Wiley and Sons.
- [3] Boots, Barry N., and Getis, Arthur (1988), *Scientific Geography Series: Point Pattern Analysis, Vol. 8*, Grant Ian Thrall, ed., Newbury Park: Sage Publications.
- [4] Diggle, Peter J. (1983), *Statistical Analysis of Spatial Point Patterns*, London: Academic Press Inc.
- [5] Cressie, Noel A. (1991), *Statistics For Spatial Data*, New York: John Wiley and Sons.
- [6] Allison, Paul D. (1995), *Survival Analysis Using the SAS(R) System: A Practical Guide*, Cary, NC: SAS Institute Inc.