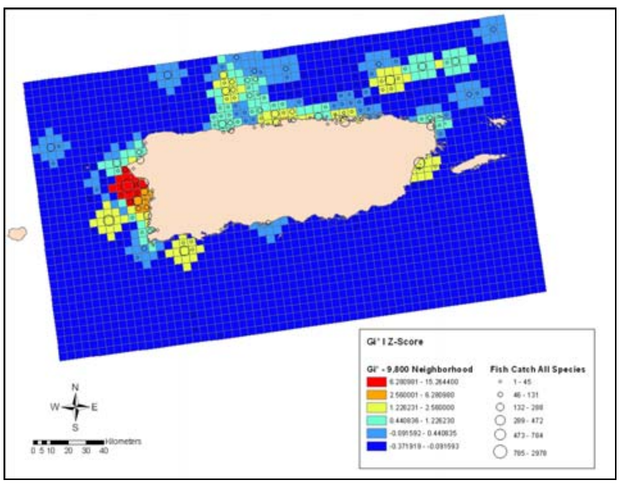
NOAA Technical Memorandum NFMS-SEFSC-541

SPATIAL STATISTICAL ANALYSIS OF THE MARINE RECREATIONAL FISHERIES STATISTICAL SURVEY (MRFSS) IN PUERTO RICO

BY

RYAN R. MORRELL



U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Southeast Fisheries Science Center
75 Virginia Beach Drive
Miami, Florida 33149

June 2006

This page intentionally left blank

NOAA Technical Memorandum NFMS-SEFSC-541

SPATIAL STATISTICAL ANALYSIS OF THE MARINE RECREATIONAL FISHERIES STATISTICAL SURVEY (MRFSS) IN PUERTO RICO

BY

RYAN R. MORRELL
Mamco, Inc.
146 12th PL SE
Vero Beach, FL 32962

U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary

National Oceanic and Atmospheric Administration
Conrad C. Lautenbacher, Jr., Undersecretary for Oceans and Atmosphere

National Marine Fisheries Service
William T. Hogarth, Assistant Administrator for Fisheries

June, 2006

NOTICE

Dr. David Carter and Brad Gentner also contributed to this report.

This report should be cited as follows:
Morrell, Ryan R. 2006. Spatial Statistical Analysis of the Marine Recreational Fisheries Statistical Survey (MRFSS) in Puerto Rico. NOAA Technical Memorandum NMFS-SEFSC-541, 61 p.

This report will be posted on the SEFSC web site at URL: http://www.sefcs.noaa.gov

Copies of this report can be obtained from:

Dr. David Carter
Southeast Fisheries Science Center
75 Virginia Beach Drive
Miami, FL 33149

or

National Technical Information Center
5825 Port Royal Road
Springfield, VA 22162
(703) 605-6000, (800) 553-6847
http://www.ntis.gov/numbers.htm

Abstract

The National Oceanic and Atmospheric Administration (NOAA) conducts the Marine Recreational Fisheries Statistics Survey (MRFSS) to develop estimates of fishing effort and catch in the territorial U.S.. During the 2003-04 study period, questions designed to elicit fishing location information were added to the MRFSS survey in Puerto Rico. The purpose of this analysis is to provide quantitative results on the spatial distribution and frequency of fish catch in the study area and calculate statistical measures of local indicators of spatial association.

The locations and densities of fish were identified using GIS. The spatial distributions of fish data were analyzed using spatial statistics in GIS. This analysis was conducted to determine if the distributions of fish in the study area occur in statistically significant patterns and if the data occur in clusters with areas of similarly high or low values of fish catch values. Tests of statistical significance were also conducted.

Table of Contents

List of Tables

List of Figures

Introduction

Problem Statement

The National Oceanic and Atmospheric Administration (NOAA) conducts the Marine Recreational Fisheries Statistics Survey (MRFSS) to develop estimates of fishing effort and catch in the territorial U.S.. During the 2003-04 study period, questions designed to elicit fishing location information were added to the MRFSS survey in Puerto Rico. The additional questions were a part of an Economic Add-On survey that included a series of maps created from a Geographic Information System (GIS) at the Southeast Fisheries Science Center in Miami, FL. Anglers were asked to identify the location of where they fished by indicating these areas on maps created in GIS.

The purpose of this analysis is to:

- conduct a literature review on spatial statistics and the aggregation of ecological data;
- provide quantitative results on the spatial distribution and frequency of fish catch in the study area;
- calculate statistical measures of local indicators of spatial association; and
- give recommendations for future study.

Prior to analyzing the MRFSS data, several general questions were identified in order to guide the analysis of the dataset. These questions are:

- Where are fish located in the waters around Puerto Rico?
- Where are specific species of fish located in the waters around Puerto Rico?
- Is it possible to create statistically relevant aggregations of fish based on the MRFSS dataset?
- Why are fish located in specific areas?

The locations and densities of fish in the waters around Puerto Rico were identified using a GIS analysis of the MRFSS data. The spatial distributions of fish data were analyzed using spatial statistics in GIS. This analysis was conducted in order to determine if the distributions of fish in the waters around Puerto Rico occur in statistically significant patterns and if the data occur in clusters with areas of similarly high or low values of fish catch values. Tests of statistical significance were also conducted when appropriate.

Based on this analysis, this MRFSS dataset does not support the creation of statistically significant aggregations or clusters based on fish location and catch values. This is due primarily to the sampling design and relatively small spatial distribution of fish across the study area. Additional data collection at varying spatial scales is required to produce meaningful aggregations of fish populations off the coast of Puerto Rico.

Literature Review

Spatial Statistics

Spatial statistics were used in the analysis of the MRFSS data. Spatial statistics vary from classical experimental statistics in several significant ways. These differences are primarily due to the non-random nature of spatial data, in this case fisheries data. As noted by Wong and Lee in their book *Statistical Analysis of Geographic Data* (2005),

> When analyzing data statistically, each observation should be independent so that its values or data are not dependent on, or tied to values of other observations in the same data set. This independence assumption is one of the most fundamental assumptions in statistical analysis. Unfortunately, it is often violated for data collected to describe events or phenomena that are spatially referenced. This is because, in many geographic events or phenomena, what happened in a location is highly correlated to what happened at its surroundings (p. 2).

The citation above relates to the MRFSS dataset because these data are not randomly distributed. Sampling efforts were systematically designed and focused on specific areas and intercept points[1]. Additionally, the measure of fish catch and location do not have an equal probability of occurring at any given position. The position of any given feature is likely to have an influence on any other feature or observation in the dataset. The location of fish off the coast of Puerto Rico is not random or spatially independent. The presence of fish is likely affected by numerous variables including time of year, currents, bottom structure, and resource availability to name a few. Therefore, this dataset violates many of the assumptions of inferential statistics and special statistical test must be used when analyzing this dataset.

The presence of spatial autocorrelation in the MRFSS must also be acknowledged when conducting statistical analysis on this dataset. Spatial autocorrelation occurs when there is the presence of any systematic pattern in the spatial distribution of a variable, in this case fish catch (Wong and Lee, 2005). Specifically, spatial autocorrelation consists of two patterns. Positive spatial autocorrelation occurs if nearby or neighboring areas are more alike. Negative autocorrelation describes patterns in which neighboring areas are unlike. Random patterns exhibit no spatial autocorrelation (Wong and Lee, 2005).

Wong and Lee (2005) further describe the impacts of spatial autocorrelation on data analysis in the following way,

> It is reasonable to suspect that the magnitude of spatial autocorrelation is not necessarily uniform over the region, but rather varies from one part of it to another. In other words, it is likely that the magnitude of spatial autocorrelation is high in some subregions but low in others. In order to capture the spatial variability of spatial autocorrelation, we must rely on another set of measures that can indicate spatial autocorrelation at the local scale. These types of measures are referred to as local indicators of spatial association (LISA) (387).

Due to the presence of spatial autocorrelation commonly found in wildlife and fisheries data, LISA statistics were used to identify clusters of grid cells with similarly high or low fish catch values. The GIS calculates a statistic for each grid cell that indicates the

---

[1] For specific details on the sampling efforts and methods associated the MRFSS, refer to MRFSS website http://www.st.nmfs.gov/st1/recreational/overview/overview.html

degree to which the nearby grid cells have similar fish catch values. These statistical values were then mapped to determine the presence or absence of clusters. Mapping the output of the LISA statistic is more insightful than simply mapping the values for the fish catch attribute (Mitchell, 2005). Mapping the output values of the LISA analysis provides statistical data to aid in identifying clusters of cells with high or low fish catch values.

Modifiable Areal Unit Problem (MAUP) and Issues of Scale

As previously noted, one of the objectives of this study was to identify clusters of fish and attempt to aggregate these clusters into statistically significant areal units. However, the aggregation of data into areal units presents several potential dangers. The primary issue is known as the modifiable areal unit problem (MAUP). When performing spatial data analysis, a primary concern is the sensitivity of statistical results to the definition of spatial units over which data are collected. In the geographical literature, this sensitivity is known as the MAUP, and results have been shown to vary with both the level and configuration of aggregation (Scanvara *et al*., 2002). According to Dugan *et al.*, "The MAUP states that changing the shape and or size of the units on which data are mapped can change the resulting correlations or statistical models generated from the data" (Dugan *et* al., 2002, p.632). Studies and research have been conducted on the MAUP and its impacts on ecological data and a conclusive solution to the MAUP is difficult to derive.

Dugan *et al*. in their article, A Balanced View of Scale in Spatial Statistical Analysis (2002), highlight the following implications of scale when evaluating ecological processes,

> In ecology, it is well know that observation scales can influence ecological inference. The characteristics of a variable's distribution depend on the area or volume over which it is measured or calculated. Specifically, inference about the population mean and variance, the strength and character of spatial autocorrelation, spatial anisotropy, patch and gap sizes, as well as multivariate relationships are all dependent on the size and shape of sampling units and the lags and extent of sampling (p. 631).

The terms ecological bias or ecological inference problem are also used in the literature when describing the problems associated with the MAUP when applied to ecological phenomena. As with the MAUP, ecological bias is based on two problems. The first problem is the aggregation bias, which is due to the grouping of individuals. The second problem is the specification bias, which is due to the different distribution of confounding variables created by grouping (Gotway and Young, 2002).

The MAUP can be addressed when designing the sampling scheme. In the case of this study, areal units of 2.5 minute cells were used to collect data from fisherman on the spatial locations of fish offshore of Puerto Rico. Dugan *et al.* indicate several factors should be acknowledged when designing a sampling methodology for ecological phenomena. Specifically, the sampling methodology should be, "…carried out using sampling units, sizes, shapes, lags and extents that are appropriate to detect the patterns or processes ones is looking for" (Dugan *et al*., 2002, p. 637). Some preliminary questions Dugan *et al.* suggest for consideration during the design of the sampling scheme include:

- What is the unit of the process?
- How big are these objects?
- How much space is affected by a process?
- What is the average distance between unit objects or processes?
- What is the spatial pattern of the distribution of these objects or processes?

To answer these questions, the spatial scale components of the sampling design have to allow detection of the pattern that may exist in the study area (Dugan *et al*., 2002, p. 637).

Since this dataset represents the first attempt at collecting spatial data associated with the MRFSS in the Southeast region, many answers to the questions outlined above were likely unknown when the sampling methodology was devised. Because of these unknown issues associated with scale and the MAUP in the MRFSS dataset, it is unlikely that the MAUP will be avoided. However, Svancara *et al.* (2002) in their study on MAUP and its implications for aggregating elk aerial survey data offer insight on the MAUP and its implications for ecological phenomena. Specifically, they suggest addressing the following three points will minimize the MAUP.

Their first point is, "meaningful results will come only from meaningful boundaries related to problem in question" (Svancara *et al.*, 2002, p.784). They conclude that by theoretically identifying units that are ecological significant, the MAUP would essentially disappear (Svancara *et al.* 2002). This point highlights the need to collect data at varying spatial scales and compare the results of the statistical analysis and aggregations at the various scales until an ecological significant unit is defined. The incorporation of secondary datasets may also assist in developing meaningful aggregations.

The second point is that researchers must determine whether the level and configuration of spatial units is appropriate for the research questions being asked (Svancara *et al.*, 2002, p. 784). The authors further state that the researchers must acknowledge if they are examining characteristics of the area or of the individuals who live there (Svancara *et al.*, 2002, p. 784). In the case of the MRFSS data, the ecological information of interest is at the individual or fish level. The implications for this study are the statistics describe the 2.5 minute cells rather than fish caught by individual anglers. The areal-level correlations are applicable to the cell and not necessarily to individual fish or anglers (Svancara *et al.*, 2002, p.784). The final point is the level and configuration of aggregation must correspond to those experienced by the population being studied (Svancera *et al.*, 2002, p. 784). This last point presents a difficult dilemma when attempting to aggregate data on a dynamic and difficult to sample population such as fish or individual anglers.

It is difficult to determine if the 2.5 minute cell size is the appropriate configuration for creating statistically significant aggregations in Puerto Rico. Since this is the first attempt at collecting and analyzing spatial data associated with the MRFSS in the Southeast, future efforts should look at collecting data at a different spatial scales, to include the discrete point locations of fish, and comparing the results to this study.

Methods

The data from the MRFSS data is contained in several data tables. The primary tables used in this analysis include the i2g_pr (type 2), i3g_pr (type 3), and i5g_pr (type 5) tables. Not all of the data contained in the tables were used in this study. The type 2 and type 3 data contain the fish catch values used in this analysis. Specifically, type 2 data were used to collect the number of species of fish caught by the anglers, but not available for inspection. Type 3 data were used to collect the number of fish caught and inspected. Type 5 data were used to collect information about where the anglers hooked the majority of the fish.

The type 5 data contains the spatial data used in this analysis. For this study, four grids were created to collect information on where fish were hooked by the fisherman sampled in the MRFSS. Each grid is composed of multiple cells and the individual cell size is 2.5 minutes, or 20.5 square kilometers. The four grids were merged into one file to simplify the analysis. All attributes of the four grids were retained when the grids were merged. A mask operation was then preformed on the merged grid to eliminate all cells that were completely contained by the boundary of Puerto Rico. The resulting grid contains only cells that are located offshore of Puerto Rico.

To create a grid for total fish catch, the following operations were completed. The type 2 and 3 data contain a unique identifier that is used to join these data to the type 5 data. Prior to joining the fish catch data from the type 2 and 3 tables to the spatial data contained in the type 5 data, several summary operations were preformed on these datasets. The type 2 and 3 data were summarized on the trip identification field (ID_CODE) and the fish catch fields (NUM_FISH and FSHINSP). This summary was necessary to create a one to one relationship between the trip identifier field, the fish catch fields and the grid location field. Once the type 2 and 3 tables were summarized to the grid cell level, they were joined to the type 5 data. The fish catch values from the type 2 and type 3 data were then calculated into one fish catch field. The resulting grid dataset contains a unique value for fish catch that is associated with a specific grid cell. This aggregation of data was necessary to perform the spatial statistical analysis.

In addition to the grid for total fish catch for all species, several grids were created that contained fish catch values for individual species. Table 1 illustrates the game fish species with the highest fish catch values. Methods similar to those used to create a grid for fish catch for all species were used to create species level grids for density mapping and statistical analysis. The primary difference in methods for the species level aggregation involved joining the type 2 and 3 tables to the species code tables in order to summarize the fish catch values at the species level.

Data loss occurred when the MRFSS data was joined to the grid. This is due to null or error values collected in the spatial identification field gridkech from the type 5 data. The primary cause for the data loss is the occurrence of null values or values of 9999 or 9998 in the gridkech field. These values do not correspond to a grid cell and cannot be joined to the grid. All fish catch data associated with these records are excluded from analysis. Table 2 illustrates the amount of data loss due to this issue.

Density maps were created for the sum of all species as well as the individual species identified in Table 2. Density maps were created using graduated symbols based on the fish catch values. The results of these maps provide a visual representation of the distribution and density of fish catch in the waters around Puerto Rico.

Local Indicators of Spatial Association

Following the creation of the density maps, the spatial statistical analysis was conducted on the grids containing data for all fish species. The local indicators of spatial association (LISA) statistics used in this study were the Local Moran's I and the Gi*. A distinction must be made between global measures and local measures of spatial autocorrelation. Global measures provide a *single value* that indicates the level of spatial autocorrelation within the fish catch variable distribution for the entire grid. Local measures provide a value for *each cell* within the fish catch variable distribution. Local indicators of spatial autocorrelation are able to identify discrete spatial patterns that may not otherwise be apparent by quantifying the spatial dependence between each cell and a surrounding area based on a defined number of cells (Holden *et al*., 2000).

Global measures of spatial association were not preformed on the MRFSS dataset due to the nature of the fisheries data. As previously noted, the location of fish in the Puerto Rican waters likely varies from subregion to subregion. A global measure that produces a single statistic for the entire dataset is not as meaningful as the outputs of the LISA statistics that produce a statistic for each cell in the grid. The global measure analyses the difference in fish catch from one location to that of the entire dataset, while the LISA measures the difference in fish catch from one cell to that of the neighboring cells.

Local Moran's I

Local Moran's I compares the value of each feature in a pair, in this case a pair of cells and their fish catch values, to the mean value for fish catch in the entire study area (Mitchell, 2005). For Local Moran's I, the fish catch values of the target feature and the adjacent feature are both compared to the mean fish catch value (Mitchell, 2005). A large positive value for Local Moran's I indicates the cell is surrounded by cells with similar fish catch values, either high or low. Conversely, a negative value for I indicates that the cell is surrounded by features with dissimilar fish catch values (Mitchell, 2005).

The Local Moran's I also calculates a Z-score. The Z-score allows for the measurement of whether each value of I is statistically significant at a given confidence level (Mitchell, 2005). A high positive Z-score for a feature indicates the surrounding values are similar values, high or low. A very negative Z-score indicates the feature is surrounded by dissimilar values, high or low (Mitchell, 2005). The Z-scores can also be classified as being above or below the expected values from a random distribution or mapped at confidence level intervals, such as 90%. The results of these maps would indicate which clusters are the most significant (Mitchell, 2005).

In his book *The ESRI Guide to GIS Analysis, Volume 2* (2005), Mitchell further explains the calculation of Local Moran's I in this way,

> The GIS first calculates the mean value (fish catch) for the attribute (cell) being analyzed. Then it calculates the difference from the mean for each neighbor and multiplies it by the weight for that neighbor. Next, it sums these products. Finally, it multiplies the sum by the ratio of the difference from the mean for the original feature's attribute value, divided by the variance (167).

The equation for Local Moran's I is

$$I_i = \frac{(x_i - \bar{x})}{s^2} \bullet \sum_j w_{ij}(x_j - \bar{x})$$

where the mean value ($\bar{x}$) is subtracted from the value of the neighbor (x_i) and the difference multiplied by the weight (w_{ij}) for the target-neighbor pair; the results for all neighbors are summed.[2] Then the sum is multiplied by the difference between the mean value ($\bar{x}$) and the target feature value (x_j), divided by the variance (s^2) (Mitchell, 2005, p.167). In this analysis, the weight value is either 0 or 1 depending on the neighborhood size. All cells outside the specified neighborhood size are assigned a weight of zero and therefore do not impact the expected value.The resulting output provides an indication of the similarity or dissimilarity of fish catch values at the local level. A large positive value for Local Moran's I indicates the cell is surrounded by cells with similar fish catch values, either high or low. Conversely, a negative value for I indicates that the cell is surrounded by features with dissimilar fish catch values (Mitchell, 167).

A Z-score is also calculated for the Local Moran's I. The Z-score formula calculates an expected Local Moran's I Value (I_i) for each cell and then compares the expected value to the calculated I_i for each cell. The Z-score for each cell is obtained by:

$$Z(I_i) = \frac{I_i - E(I_i)}{\sqrt{Var(I_i)}}$$

[2] Refer to the spatial neighborhoods section on page 10 for the application of the weight value in this analysis.

for each feature, the expected (I_i) value is subtracted from the observed $E(I_i)$ and the difference is divided by the square root of the variance (Mitchell, 2005, p.170).

The expected $E(I_i)$ is calculated by taking the negative of the sum of the weights and dividing by the number of features, minus 1 (Mitchell, 2005).

$$E(I_i) = \frac{-\sum_j w_{ij}}{n - 1}$$

Where $E(I_i)$ is the expected Local I value and the negative sum of the weights for all feature pairs (w_{ij}) is divided by the number or features (n) minus one (Mitchell, 2005 p. 170). In this analysis, the weight value is either 0 or 1 depending on the neighborhood size. All cells outside the specified neighborhood size are assigned a weight of zero and therefore do not impact the expected value. However, the literature notes that the Z-score can produce misleading results when used with local measures of spatial association, such as the Local Moran's I (Mitchell, 2005). This issue arises because the Z-score test assumes a random distribution and independence between features. However, as previously noted, the occurrence of fish in individual cells is not randomly distributed or independent. The independence test in this case is violated because fish catch values for some adjacent cells appear to be similar. In fact, the detection of this pattern of clusters with similar fish catch values is a primary goal of this study. The specific impacts from this issue are not easy to determine and do not preclude the use of the Z-score for testing statistical significance. Nevertheless, the results may be suspect because of the violation of independence between values.

<u>Gi*</u>

For the Gi* statistic, the GIS sums the fish catch values of the neighbors and divides by the sum of the values of all the features in the study area (Mitchell, 2005). The number of features in the study area is determined by the extent of the neighborhood size. A group of cells with high Gi* values indicates a cluster of features with high fish catch values. A group of features with low Gi* values indicates a cluster of features with low fish catch values. A Gi* value near 0 indicates there is no concentration of either high or low values surrounding the target feature (Mitchell, 2005).

The equation for Gi* is

$$Gi^*(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}$$

where Gi* is calculated for a feature (i), at a distance (d). The value of each neighbor (x) is multiplied by the weight for the target-neighbor pair (w_{ij}), and the results are summed and then divided by the sum of the values of all neighbors (x_i) that is, all features in the data set (Mitchell, 2005, p. 176). In this analysis, the weight value is either 0 or 1 depending on the neighborhood size. All cells outside the specified neighborhood size are assigned a weight of zero and therefore do not impact the Gi* value.

A Z-score is also calculated for the Gi* analysis. The Z-score is calculated by subtracting the expected Gi* value for the cell, given a random distribution, from the calculated Gi* value. The difference is then divided by the square root of the variance for all features in the study area (Mitchell, 2005).

The equation for the Gi* Z-score is

$$Z\,(G_i*) = \frac{G_i* - E(G_i*)}{\sqrt{Var(G_i*)}}$$

where the expected Gi* value is subtracted from the observed Gi* and the difference divided by the square root of the variance (Mitchell, 2005, p.178).

The equation for the expected Gi* is

$$E\,(Gi*_i) = \frac{-\sum_j w_{ij}(d)}{n-1}$$

where the weights at (w_{ij}) at a distance are summed and divided by the number of features (n), minus one (Mitchell, 2005, p. 178). In this analysis, the weight value is either 0 or 1 depending on the neighborhood size. All cells outside the specified neighborhood size are assigned a weight of zero and therefore do not impact the expected value.

A high Z-score for a cell indicates its neighbors have a high fish catch value and a low Z-score indicates a cell's neighbors have a low fish catch value. A Z-score near 0 indicates a lack of concentration of similar values. To determine if the Z-score is statistically significant, it can be compared to the range of values for a given confidence level (Mitchell, 2005). As with the Z-score for Local Moran's I, the Z-score for Gi* is also suspect when the independence test is violated because the calculation for the expected value assumes a random distribution.

Spatial Neighborhoods

The notion of the neighborhood of surrounding values must also be discussed to better understand the LISA statistics. The LISA statistics used in this analysis evaluate both the difference between the values of the features (fish catch) and the spatial relationship between the features (distance) (Mitchell, 2005). As described by Mitchell (2005), during the analysis the GIS compares the value of a target cell to the fish catch values of the neighboring features. The GIS then moves to the next feature until the analysis has been completed on all features in the defined neighborhood. The user determines the size of the defined neighborhood and variations in the neighborhood size will impact the statistical output.

The large number of cells in the dataset with no fish catch values will impact the output of the LISA statistics. Large areas of contiguous cells with no fish catch values likely represent data outliers or under sampled areas. Although not all cells with no fish catch values represent data outliers, the large geographic extent of the study area and the systematic nature of the sampling techniques utilized are likely to produce over and under sampled areas. When using LISA statistics to identify clusters of features with similar values, large areas of features with no data (zero fish catch value) may skew the results and indicate either clusters of similar or dissimilar data when in reality these areas may represent under sampled areas with no observed data.

The sampling design for this study utilized a cell size of 2.5 minutes or approximately 20.5 square kilometers. It is important to note that the points derived from associating the gridkech field from the type 5 data to the fish catch values in the type 2 and 3 data represent continuous data summarizing the number of fish caught in the entire 2.5 minute cell. Although the data are represented as points for the spatial analysis, the data are not discrete. In other words, the point data representing the centroid of each cell do not represent the discrete location of fish, but are a summary of the fish caught in the cell. The fish catch data are interval data and represent quantities of fish caught in a cell.

For all measures of local spatial association, the definition of neighborhoods for adjacent cells was distance-based. In the GIS, both the Local Moran's I and Gi* operation require the user to specify the neighborhood size. The distance between centroids in the MRFSS data is approximately 4,900 meters. Therefore, equal intervals of 4,900 meters were used as the threshold or cutoff distance for determining the extent of the neighborhood for the LISA statistics. Four cutoff distances were used to detect the impact on the LISA statistics by varying the neighborhood size. The cutoff distance intervals are summarized in Table 3.

For each interval, the Local Moran's I and Gi* values were calculated in ArcGIS. Because LISA statistics measure the spatial relationship between cells, a spatial weight is assigned to each cell and its neighbors. The weight value is calculated using the distance between the features (Mitchell, 2005 p. 137). In this case, the distance is the distance between the centroids of each cell. During the analysis, all neighbor points within the cutoff distance are assigned a weight of 1; all neighbors farther than cutoff distance are assigned a weight of 0.

Results

Prior to conducting the statistical analysis, the fish catch data was displayed thematically in effort to visualize the location and catch of fish data obtained from the MRFSS. Figure 1 depicts the total fish catch density values for the entire survey area and indicates clusters of relatively high fish catch values on the north and west coasts of Puerto Rico. Very few data points exist for the southern portion of Puerto Rico, indicating this area was likely under sampled. The cells with the highest fish catch values are located along the west coast of the island. Of the 1,703 grid cells in the survey area, 195, or 11% percent of the cells in the study area, have a fish catch value greater than zero.

Species-specific densities for the game species with the highest catch values relative to the entire dataset are depicted in Figures 2 through 6. Table 4 summarizes the species-specific data by fish catch and spatial extent.

The data indicate a relatively small amount of grid cells with fish catch data when compared to the entire study area. As previously noted, this is likely due to the sampling design and the aggregation of point data to areal units. Mitchell (2005) indicates that the output of LISA statistics is suspect when fewer than 30 features are analyzed. Based on the data in Table 4, only one species, Dolphin, has a spatial distribution significantly larger than 50 (number of cells with fish catch greater than 0). Therefore, LISA statistics were computed only for dolphin as well as for the entire dataset.

Figures 7-10 depict the results of the Local Moran's I analysis used to identify clusters of cells with similar fish catch values. This analysis was preformed on the entire dataset. For all figures, a high positive value (red to orange) indicates that the cell is surrounded by cells with similar fish catch values, either high or low. A negative value (dark blue) indicates the cell is surrounded by cells with dissimilar values, either high of low. In Figures 7-10, total fish catch values are displayed thematically with graduated circles in conjunction with the Local Moran's I score to aide in identifying cells with similar and dissimilar fish catch values.

Although a high I value indicates a cluster of cells with similar fish catch values, low (negative) I values may also aide in identifying statistically significant areas of fish occurrence. If a grid cell has a high fish catch value, but is surrounded by cells with low fish catch values, the resulting I value will be low. If the cells with low or no fish catch data are representative of that area and do not represent outliers or under sampled areas, then the cell with a low I value may indicate an area with a statistically significantly high fish catch value in comparison to adjacent areas.

Figures 11-14 also depict the Local Moran's I value but are symbolized in a different manner than Figures 7-10. The values depicted in the Figures 11-14 represent standard deviations from the mean I value. Cells with either high or low deviations from the mean value may indicate statistically significant areas of fish catch. The thematic display based on standard deviations is another method of symbolizing the data and is insightful in identifying cells significantly different than the mean I value.

Tables 5 and 6 compare the descriptive statistics calculated on the Local Moran's I value for the four neighborhood sizes. Several trends are visible by reviewing Tables 5 and 6. The mean, range, and standard deviation for I increase as the neighborhood size increases. This is reasonable because the inclusion of more neighbors (cells) in the analysis is likely to increase the mean value for fish catch in a neighborhood. Also, the maximum value for I increases dramatically with an increase in neighborhood size. This is expected because an increase in neighborhood size results in a potential increase in the fish catch values due to the inclusion of more cells in the analysis. However, it is difficult to select a best-fit neighborhood distance.

Figures 15 through 18 depict the Z-scores for the Local Moran's I analysis. The Z-scores are mapped at either a 95% or 99% confidence interval (+/- 1.96 or +/- 2.56). In the case of spatial data, the Z-score measures the likelihood that the similarity between a feature and its neighbors is not due to chance (Mitchell, 2005). Ideally, a Z-score value higher than 1.96 indicates that cell's fish catch value is statistically similar to its neighbors at a confidence level of 0.05. Conversely, a Z-score value lower than –1.96 is statistically dissimilar to its neighbors at a confidence level of 0.05. However, as previously noted, the literature indicates that the Z-score can produce misleading results when used with local measures of spatial association, such as the Local Moran's I (Mitchell, 2005). This issue arises because the Z-score test assumes a random distribution and independence between features. However, the occurrence of fish in individual grid cells is not randomly distributed or independent. Although this issue cannot be overlooked, there appears to be a similar pattern of high I values and Z-score values in Figures 7-19, providing at least a qualitative indicator that there are several areas with clusters of cells with similarly high of low fish catch values.

By comparing the figures presented in the following pages, several patterns are apparent. Relatively high I values and associated Z-scores are consistently found along the western shoreline, indicating the cells in this area with high I values are surrounded by features with similarly high fish catch values. This area appears to contain the highest concentration of cells with similarly high fish catch values. Multiple cells in this area have fish catch values higher than 300, the likely reason for high I values.

Another pattern is visible off the northern coast. As the neighborhood size increases, the number of clusters with relatively high I values and high Z-scores decreases. This may be due to the relatively large number of cells with relatively low fish catch values in this area. Also, by increasing the neighborhood size, a larger number of cells with low fish catch values are included in the analysis. It is possible that in this region the scale of spatial autocorrelation is more localized than along the western coast. Recall that spatial autocorrelation refers to the presence of any systematic pattern in the spatial distribution of a variable, in this case fish catch (Wong and Lee, 2005). In other words, the influence of fish catch values on adjacent cells may be greater in both scale and extent along the western coast than the northern coast. Consistently low I values are found in several areas. The cells with the low I values generally have a high fish catch value and are surrounded by cells with low fish catch. This may indicate a hot spot for fish catch when compared to the neighboring areas.

Figures 15 through 18 depict the results of the Gi* statistic. The resulting output from the Gi* analysis in ArcGIS are Z-scores. The results depicted in Figures 15 through 18 are Z-scores. A high value (orange to red) indicates a cluster of cells with high fish catch values. A low value (dark blue) indicates a group of cells with low fish catch values. A value near zero indicates there is no concentration of high or low fish catch values. This occurs when adjacent values are near the mean, or when the target cell is surrounded by both high and low values (Mitchell, 2005). In Figures 15-18 the data is mapped at the 99% confidence intervals for the neighborhood sizes of 4,900 m (1 centroid) and 9,800 m (2 centroids) and the 90% confidence intervals for the neighborhood sizes of 14,700 m (3 centroids) and 19,800 m (4 centroids). It should be noted that the literature indicates that the Z-score can produce misleading results when used with local measures of spatial association, such as Gi* (Mitchell, 2005).

Mitchell (2005) indicates that when the extent of the neighborhood is large, there are likely to be a few large clusters and with a smaller neighborhood extent a larger number of small clusters will be observed. This appears to be the case in the analysis. Figures 23-34 depict the results of the Local Moran's I and Gi* analysis for Dolphin only.

Similar patterns in the I values and Z-scores are observed with both the all species analysis and the dolphin only analysis. The mean value and range of I increases with an increase in neighborhood size. As the neighborhood size increases, the cells with high I values off the northern coast are increasingly surrounded by cells with lower I values. This trend may indicate that the scale of spatial autocorrelation is more localized in this area.

There are numerous cells located off the northern coast where dolphin were repeatedly sampled and observed. Because of this, the issue of under sampling may be less of a concern when analyzing the subset of dolphin data. Clusters of high I values are consistently found off the northern coast of Puerto Rico. Additionally, there are a few cells with relatively low I values that also have high fish catch values, indicating these cells are significantly different than their neighbors.

Regardless of the neighborhood size, clusters of cells with Z-score values greater than 1.96 or 2.56 are repeatedly observed located off the northern coast. Z-score values at this level indicate a statistically significant cluster of cells with high dolphin catch values at either the 95% or 99% confidence level. These high Z-score values indicate that the similarity between the dolphin catch values for these cells is not due simply to chance.

It is difficult to draw definitive conclusions on this dataset due to the small sample size and MAUP issues associated with aggregating point data to the 2.5 minute cell. However, regardless of the neighborhood size, several cells off the northern coast have consistently high I values, Z-Scores and fish catch values. This indicates that the presence of dolphin in these areas are statistically more significant than in the surrounding areas.

Conclusions

The primary result from this study is an exploratory data analysis of the MRFSS data using GIS and spatial statistics. At the onset of this study, several questions were outlined in effort to describe the MRFSS dataset and provide direction for the statistical analysis. The questions are presented below and are followed by answers, potential solutions to the questions, or recommendations for further study.

1. Where are fish located in the waters around Puerto Rico?

Figures 1 through 6 provide a thematic display of catch and location for all fish species as well as game fish species with relatively high catch values. Based on the data for all species, a relatively high area of fish catch is located off the central portion of the western coast. Also, numerous cells with relatively low fish catch values are located off the northern coast. The mean value for fish catch off the northern coast is 40, compared to a mean value of 157 for the western coast.

There are many factors that may contribute to that fact that only a relatively few number of sampling units (11%) have fish catch values greater than one. It is likely due to sampling design and the difficulties associated with collecting spatial information on the location of fish populations. One potential source of error associated with the sampling design is the cell size. The spatial scale component (cell size) of the sampling design has to allow detection of the pattern that may exist in the study area (Svancara *et al.*, 2002). It is possible that the sampling unit, in this case 2.5 minute cells, are not the appropriate size to detect clusters of fish in the waters off Puerto Rico. This issue revolves around the problems associated with the MAUP. Namely, different inferences may be obtained when the same set of data is grouped into different size areal units (Gotway and Young, 2002). However, this study is the first attempt to collect and aggregate spatial data associated with the MRFSS in the Southeast and it is difficult to analyze the scale problem without comparing this data to data collected at varying areal units.

Several potential sources or error are injected into the sample based on the selected sampling methodology. The first being that the data associated with the gridkech field in the type 5 data is not the exact location for all fish caught on the trip, but only the location where the majority of the fish caught were hooked. Second, the selection of a specific cell by fishermen for the gridkech data is not a precise measure of fish location. Additionally, the gridkech data is an aggregation of point data to an areal unit. As previously mentioned, the point data representing the centroid of each cell do not represent the discrete location of fish, but are a summary of the fish caught in the sample cell. Finally, fish catch values from the type 2 data are based on data reported by fisherman but not observed by the researchers. Although this method of sampling is cost-effective, the nature of the survey does interject sources of error that can impact the statistics associated with this analysis.

2. Where are specific species of fish located in the waters around Puerto Rico?

Figures 2-6 depict fish catch and location for the following five game fish species: dolphin; wahoo; snook; lane snapper; and yellow tail snapper. Of these species, only two, wahoo and dolphin, are recorded at 30 or more unique cells in the study area. This is a relatively low spatial distribution and may be due to some of the issues associated with the sampling methodology previously detailed. Mitchell (2005) notes LISA statistics with less than 30 features are suspect due to the impact of outliers on the distribution of values. It is recommended that additional species-specific data be collected prior to performing LISA statistics at the species level.

Of the five species, dolphin has the largest sample size and widest spatial distribution. Several cells off the central north coast and northwest coast have relatively high catch values greater than 150. The LISA statistic results for Dolphin indicate there are several clusters of cells with high fish catch values and the Z-score tests indicate the cells with high I values are statistically significant. Based on this analysis, it is likely that the clusters of cells with high dolphin catch off the northern coast are not due to chance.

3. Are there patterns or clusters of fish in the waters around Puerto Rico?

LISA statistics were calculated to determine clusters of cells with high and low fish catch values. Figures 7 through 22 depict the results of these analyses. The LISA statistics used in this study are generally interpreted in the following manner. High values indicate clusters of cells with similar fish catch values, either high or low. Low values indicate clusters of cells with dissimilar fish catch values, either high or low. By viewing the values of the LISA statistic, the associated Z-score and the fish catch values, it is possible to identify statistically significant clusters of cells with high fish catch values.

By reviewing Figures 7 through 22, several areas consistently have high Local Moran's I values, Z-scores and fish catch values, regardless of the neighborhood size. Areas with this association of values indicate statistically significant areas of fish catch. A high Z-score value measured at a specified confidence interval indicates it is likely that these clusters of cells do not occur due to chance. Several clusters off the north and west coast of Puerto Rico contain this association of values. For Dolphin, statistically significant clusters of cells with high fish catch are located off the northern coast. However, the Z-scores associated with local statistics are at times suspect and the issues associated with the MAUP and the aggregation of the data to the 2.5 minute cells must be considered prior to making management decisions.

4. Is it possible to create statistically relevant aggregations of fish based on the MRFSS data?

The short answer is not in its present form. However, several areas have relatively high fish catch values and consistently high LISA values, indicating clusters of cells with similarly high fish catch values. As noted throughout this paper, it is difficult to aggregate catch values for wildlife populations into statistically significant clusters. This study represents the first effort at collecting spatial data in the Southeast region associated with the MRFSS. The analysis and results in this paper are exploratory in nature and provide NOAA with a background on the issues associated with aggregating ecological data as well as an application of spatial statistics to fisheries data.

Several issues arise when attempting to aggregate these data into statistically significant clusters. The first concerns the sampling methodology. In the open water, the location where a fish is caught is a discrete location. However, in these data, these discrete locations are summarized to a grid cell level of 2.5 minutes in area. The presence of fish in specific areas in the ocean is due to a variety of factors and these boundaries may not be appropriately sized to detect ecological thresholds or significant variations in fish catch. The preferred solution to this issue is to collect data at the discrete (point) locations where fish are caught. The other option is to aggregate the data at a different cell size and compare the results to this effort.

The second issue associated with this analysis relates to the neighborhood size used in the calculation of the spatial statistics. The LISA statistics use a distance function to determine local variation in fish catch. It is difficult to determine the distance threshold where the influence of a fish catch value at one location no longer influences fish catch in neighboring cells. It is likely that this distance varies across the study area and is influenced by variables such as temperature, bathometry, or bottom structure. This analysis was conducted with varying neighborhood distances. Generally, as the neighborhood size increased, the number of clusters decreased. However, it is very difficult to determine the best-fit value for neighborhood size, especially since the data are not true point features, but are summarized to an arbitrary areal unit of 2.5 minutes. This analysis does provide a foundation for measuring the impact of neighborhood size on local statistics.

5. Why are fish located in these areas?

Many factors influence fish occurrence in specific areas. It is likely high fish catch values are positively correlated with the occurrence of optimal habitat. Due to the relatively low spatial distribution of fish occurrence and catch at the species level, the majority of this analysis was conducted on the entire dataset. However, habitat preferences vary by species. Therefore, it is suggested that indicator variables for prime fish habitat by species are identified and spatially referenced for the study area. The incorporation of explanatory variables, such as indicators of optimal fish habitat, will enhance this analysis and aide in identifying ecologically significant spatial units for aggregation. Additional data collection at the species level will also produce more accurate values for the spatial statistics.

Future Study

Several modifications to the survey methodology and analysis may improve the ability to statistically identify areas with high fish catch. One approach for future study could evaluate the statistical relationship between fish catch and independent variables associated with fish habitat when the data is aggregated at varying spatial scales. In order to develop this type of aggregation, secondary data on fish populations such as habitat preferences, environmental conditions, seasonality, and other variables need to be incorporated into the analysis. The identification of indicator variables for prime fish habitat by species are identified and spatially referenced for the study area. The incorporation of explanatory variables, such as indicators of optimal fish habitat, will enhance this analysis and aide in identifying ecologically significant spatial units for aggregation. Additional data collection at the species level will also produce more accurate values for the spatial statistics.

Select variables in the Type 1 MRFSS data may also be used as independent or indicator variables to estimate catch rates associated with specific angler characteristics. For example, data such as number of days fished, time of year, boat length, etc. may be used to refine the analysis. Type 1 data for cells with high catch values may be analyzed to indicate or predict angler characteristics that are associated with high catch values. Additionally, the spatial distribution and location of cells associated with high catch values can be summarized based on Type 1 data. Hot spots or clusters of high fish catch values may be discernable based on analyzing angler characteristics. Regression analysis with angler characteristics as independent variables may also be completed in effort to predict fish catch. If the regression analysis is statistically significant, subsequent GIS analysis could focus on the spatial location of cells with statistically significant independent variables associated with predicting fish catch.

One study in the literature utilized Pearson correlation coefficients and multiple regression to evaluate the relationship between biological variables at varying levels of aggregation in effort to determine the impacts of the MAUP on various aggregations. Aggregating data based on biological configurations may produce more reliable estimates and may be more effective for policy decisions (Svancara, *et al.,* 2002).

The collection of fish occurrence data at discrete locations would also improve the spatial statistical results. Analyzing the exact location of fish occurrence would eliminate the arbitrary aggregation of all data to the 2.5 minute cell. GPS data collection devices located on fishing vessels would facilitate this type of data collection. If GPS data collection is not practical, than collecting data at different spatial scales and then comparing the results to this study would also prove useful in identifying statistically significant clusters of fish.

Geostatistcal analysis of the MRFSS spatial data is another potential application. Geostatistics involves predicting values at unknown locations based on values at known locations using kriging techniques in GIS. If accurate measures of fish location and catch can be obtained through sampling, than these data could then be used to predict fish occurrence and catch in unknown areas.

In general, the application of spatial statistics to fisheries data is a relatively recently and expanding field of study. During the course of this study, locating similar research proved elusive, although these efforts were not exhaustive. The implications of the MAUP on the management of wildlife and fisheries are increasingly found in the literature. This study is a first effort in addressing the difficulties of aggregating fisheries data and further study based on these recommendations will minimize the effect of the MAUP and improve NOAA’s foundation for managing fish populations.

Table 1: Game Fish Catch Summary

Species	Reported (Type 2)	Inspected (Type 3)	Total
Dolphin	108	2,454	2,562
Lane Snapper	92	252	344
Snook	60	130	190
Wahoo	26	589	615
Yellowtail Snapper	63	291	354

Table 2: Records Lost When Joined to Spatial Grid

Data Source	Total Records	Invalid Spatial Records	Records After Join to Grid	Percent Decrease in Total Records
Type 2	1,793	441	1,352	25%
Type 3	13,284	582	12,702	4%

Table 3: Neighborhood Size Intervals

Interval	Distance (m)
1	4,900
2	9,800
3	14,700
4	19,600

Table 4: Fish Catch and Spatial Extent

Species	Reported (Type 2)	Inspected (Type 3)	Total	Percent of All Fish	Spatial Extent (number of cells with fish catch > 0)	Percent of all cells in Study Area
Dolphin	108	2,454	2,562	17%	83	4.70%
Lane Snapper	92	252	344	2%	15	0.85%
Snook	60	130	190	1%	12	0.68%
Wahoo	26	589	615	4%	42	2.38%
Yellowtail Snapper	63	291	354	2%	21	1.19%
All Species	1,793	13,284	15,077	100%	195	11.4%

Table 5: Summary of Local Moran's I Values at Varying Intervals for All Species

Statistic	4,900 m	9,800 m	14,700 m	19,800 m
Mean	0.055374614	0.585658831	1.174095544	1.41108987
Standard Error	0.021013861	0.307575418	0.514518631	0.55550999
Median	0.034857299	0.095857501	0.244001001	0.34772
Mode	0.034857299	0.104571998	0.313715011	0.52285898
Standard Deviation	0.882832299	12.92182866	21.61590686	23.3380318
Sample Variance	0.779392868	166.9736558	467.2474293	544.66373
Kurtosis	494.1702568	852.1904281	825.526863	798.14339
Skewness	19.09403523	28.96440587	27.38375384	26.8033885
Range	32.3897295	401.6465368	748.4984169	795.500006
Minimum	-9.137330055	-8.678549767	-25.65539932	-23.982
Maximum	23.25239944	392.9679871	722.8430176	771.518005
Confidence Level(95.0%)	0.041214636	0.603249862	1.009129062	1.08952571

Table 6: Summary of Local Moran's I Values at Varying Intervals for Dolphin

Statistic	4,900 m	9,800 m	14,700 m	19,800 m
Mean	0.253683523	0.664927739	1.470899868	1.99463302
Standard Error	0.118047399	0.275561421	0.518617496	0.647767255
Median	0.0326763	0.0898597	0.228734	0.372634
Mode	0.0326763	0.0980288	0.294086	0.490144
Standard Deviation	4.941098086	11.57685969	21.78810797	27.21393515
Sample Variance	24.4144503	134.0236803	474.7216488	740.5982662
Kurtosis	765.041801	616.1239507	359.0758061	300.2170258
Skewness	26.90398432	23.15595181	17.93724756	16.3321063
Range	155.02112	384.6866	612.8552	731.411
Minimum	-5.01812	-25.0906	-61.7292	-107.68
Maximum	150.003	359.596	551.126	623.731
Confidence Level (95.0%)	0.231528817	0.54046058	1.017168195	1.270470539

Figure 1: Fish Catch for All Species

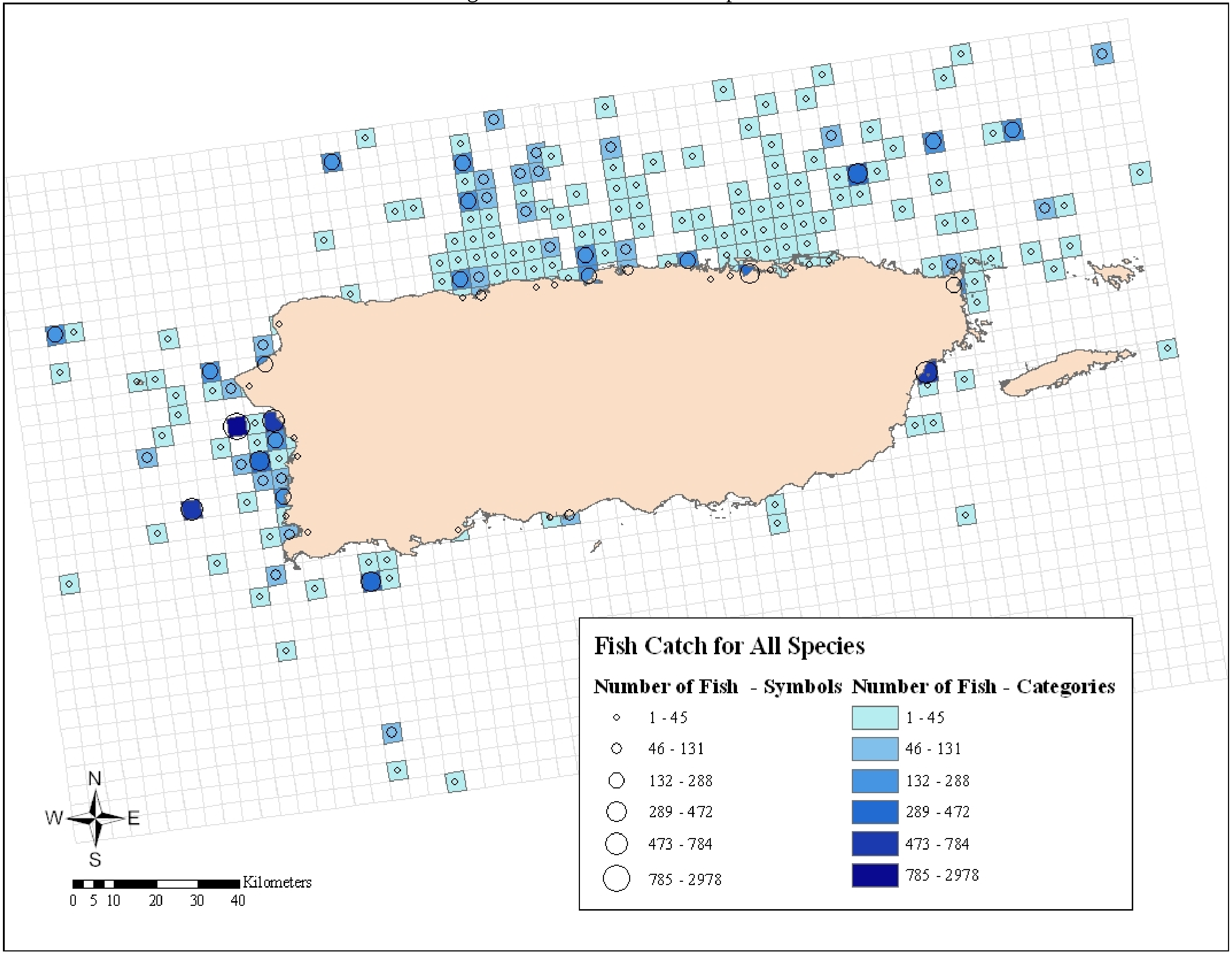


Figure 2: Fish Catch for Lane Snapper

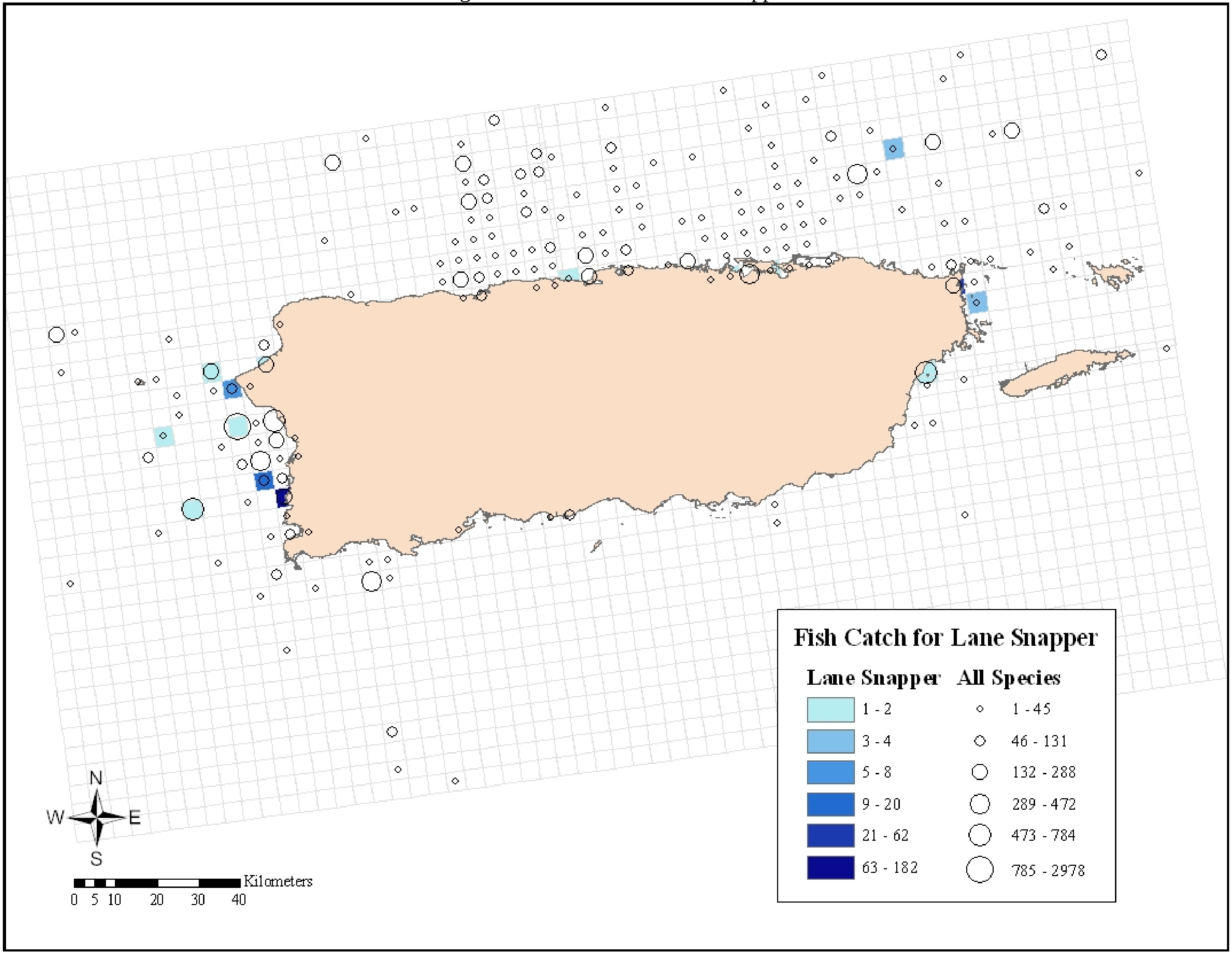


Figure 3: Fish Catch for Wahoo

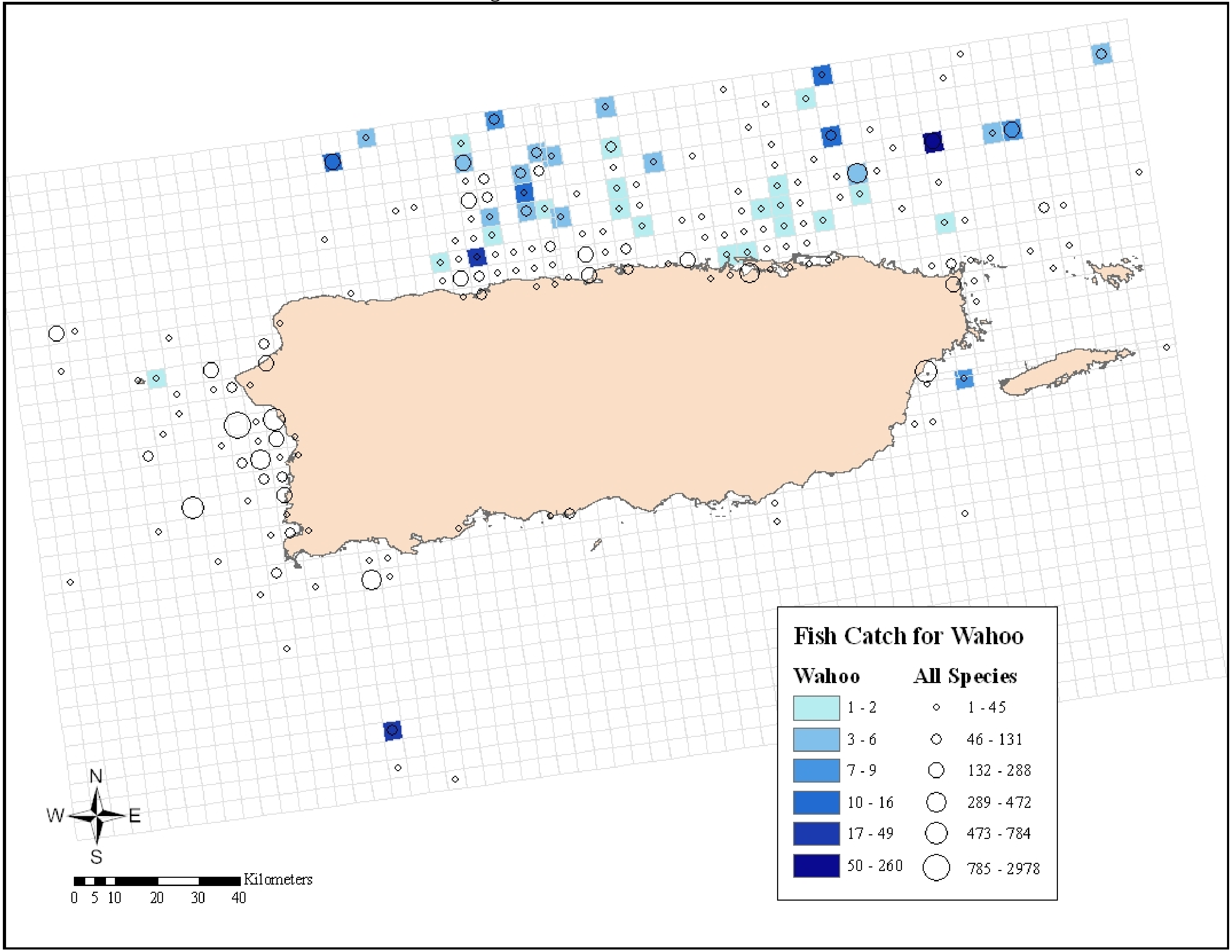


Figure 4: Fish Catch for Snook

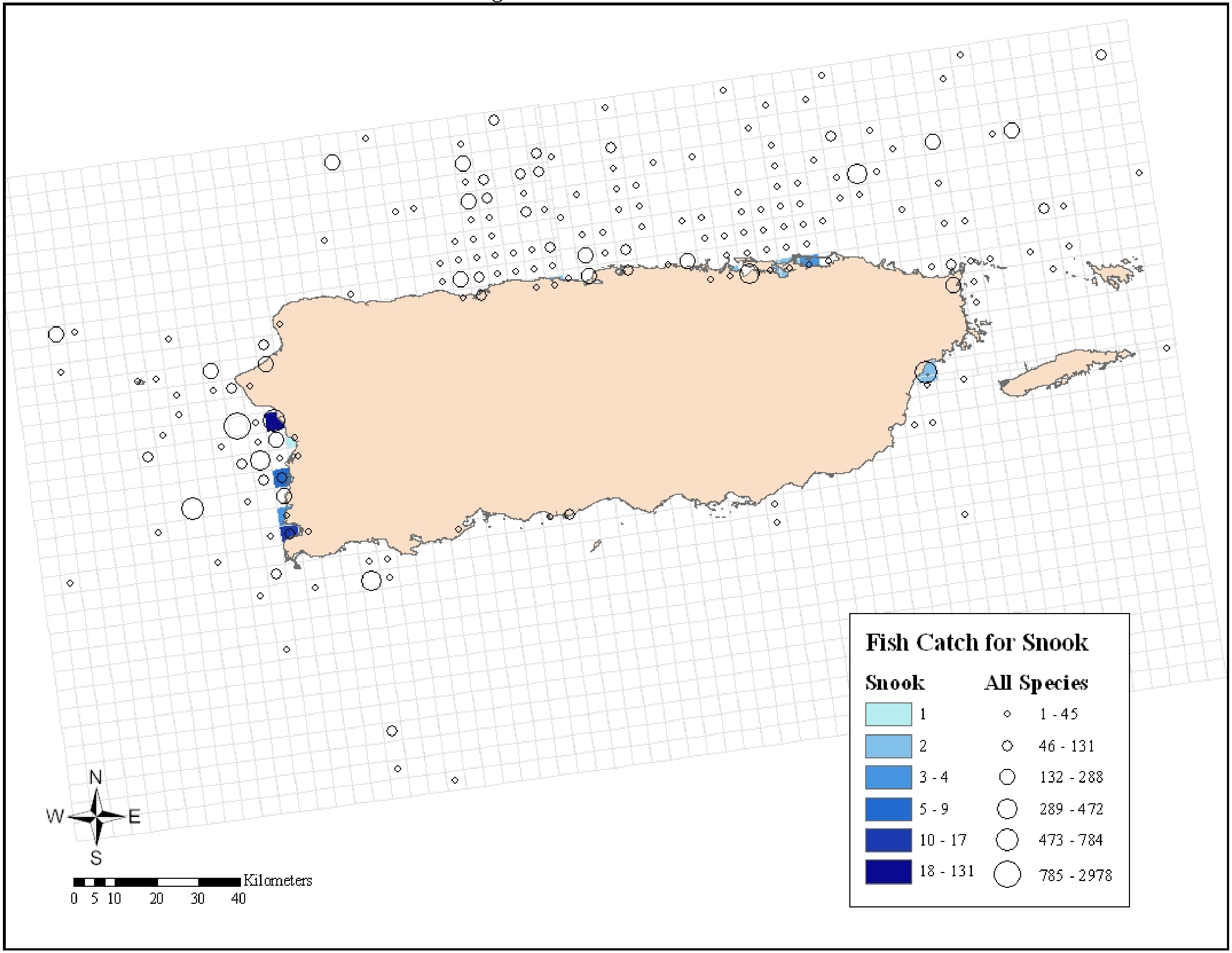


Figure 5: Fish Catch for Yellowtail Snapper

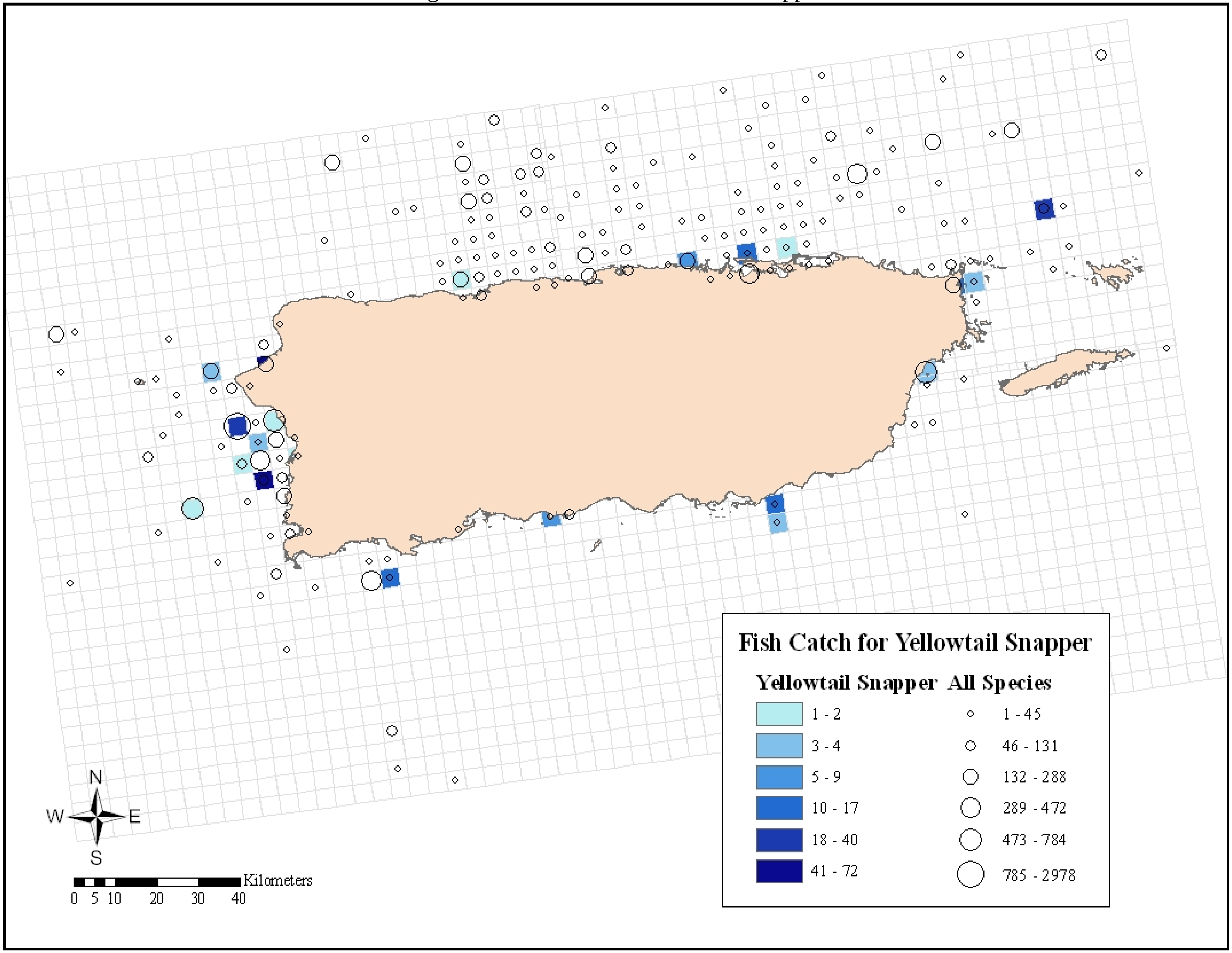


Figure 6: Fish Catch for Dolphin

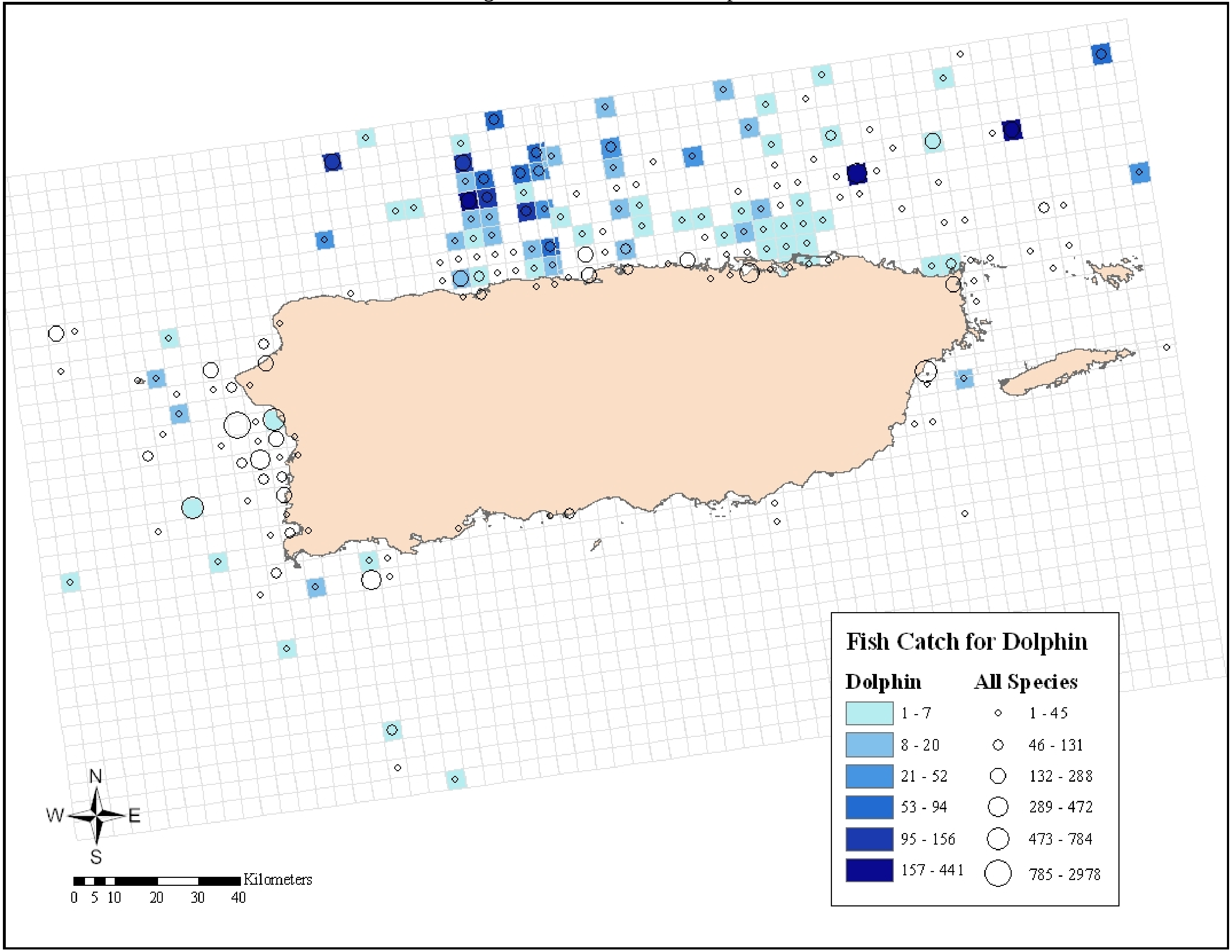


Figure 7: Local Moran's I with 4,900 m Neighborhood (1 centroid)

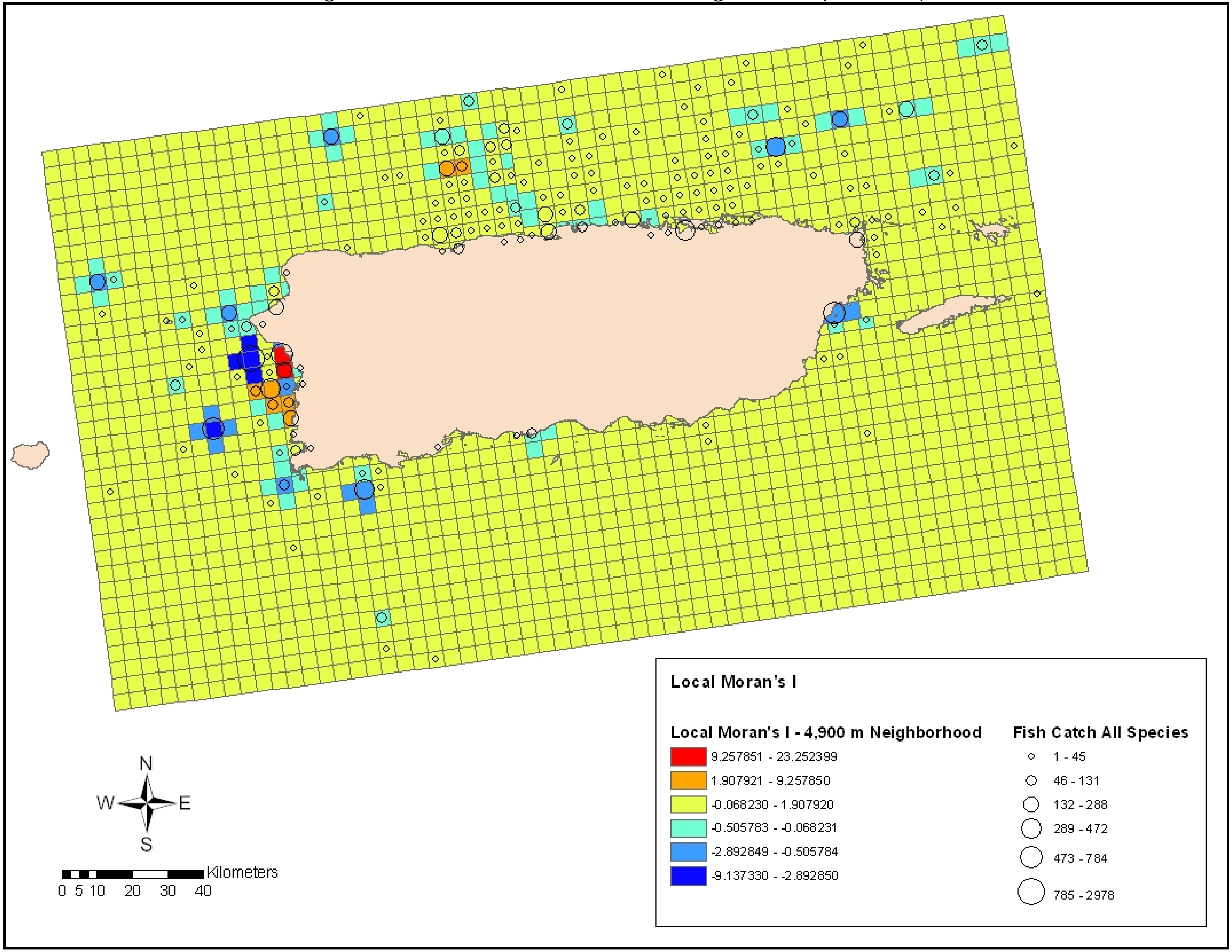


Figure 8: Local Moran's I with 9,800 m Neighborhood (2 centroids)

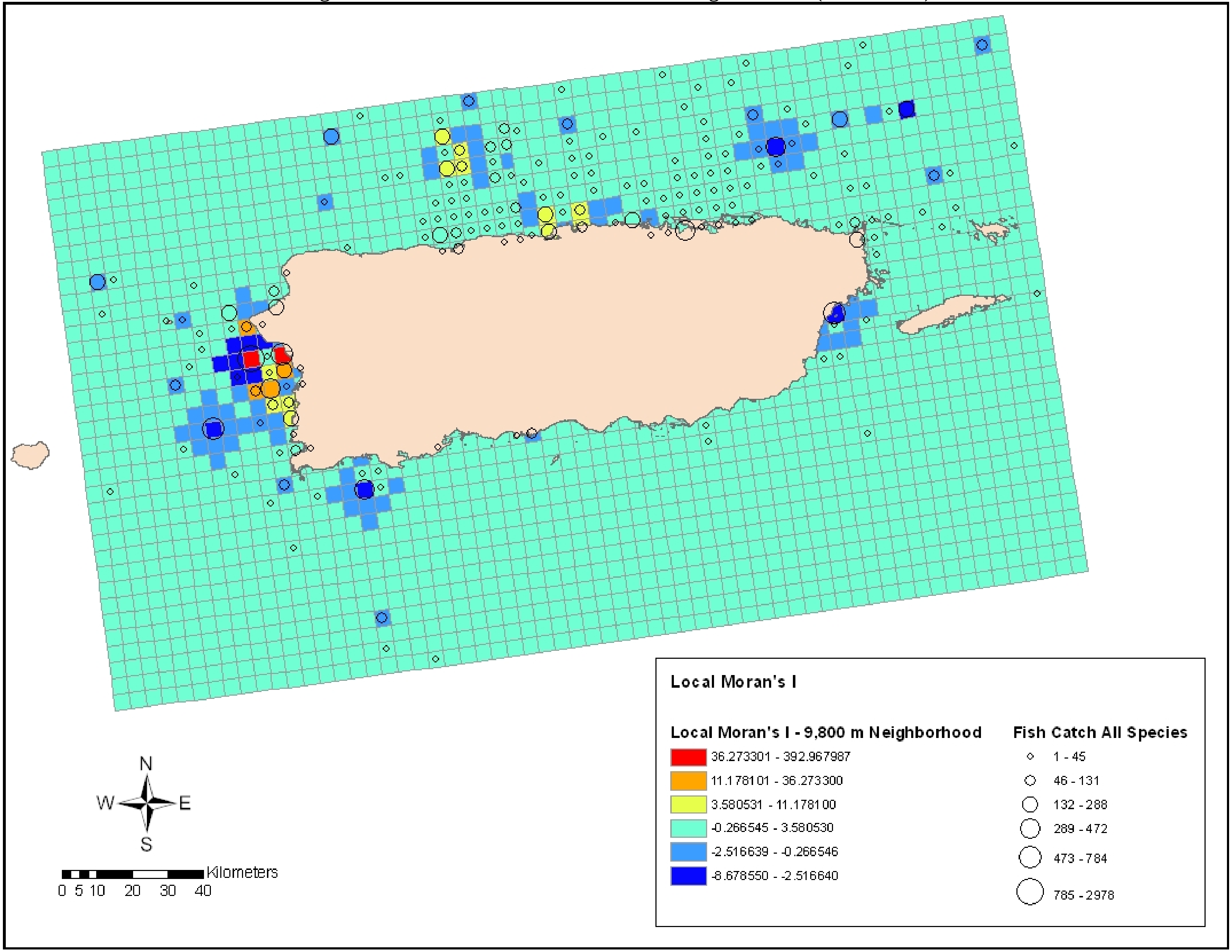


Figure 9: Local Moran's I with 14,700 m Neighborhood (3 centroids)

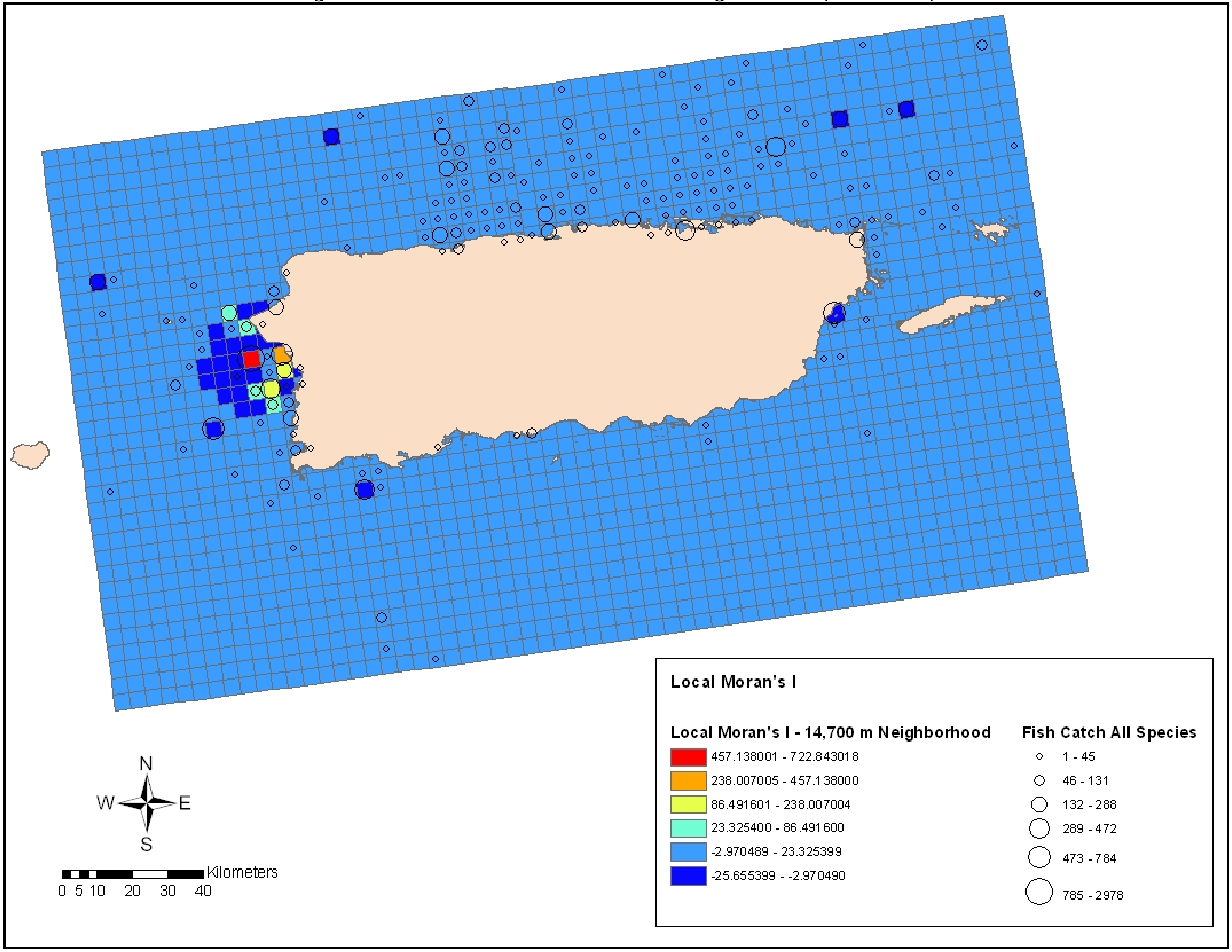


Figure 10: Local Moran's I with 19,800 m Neighborhood (4 centroids)

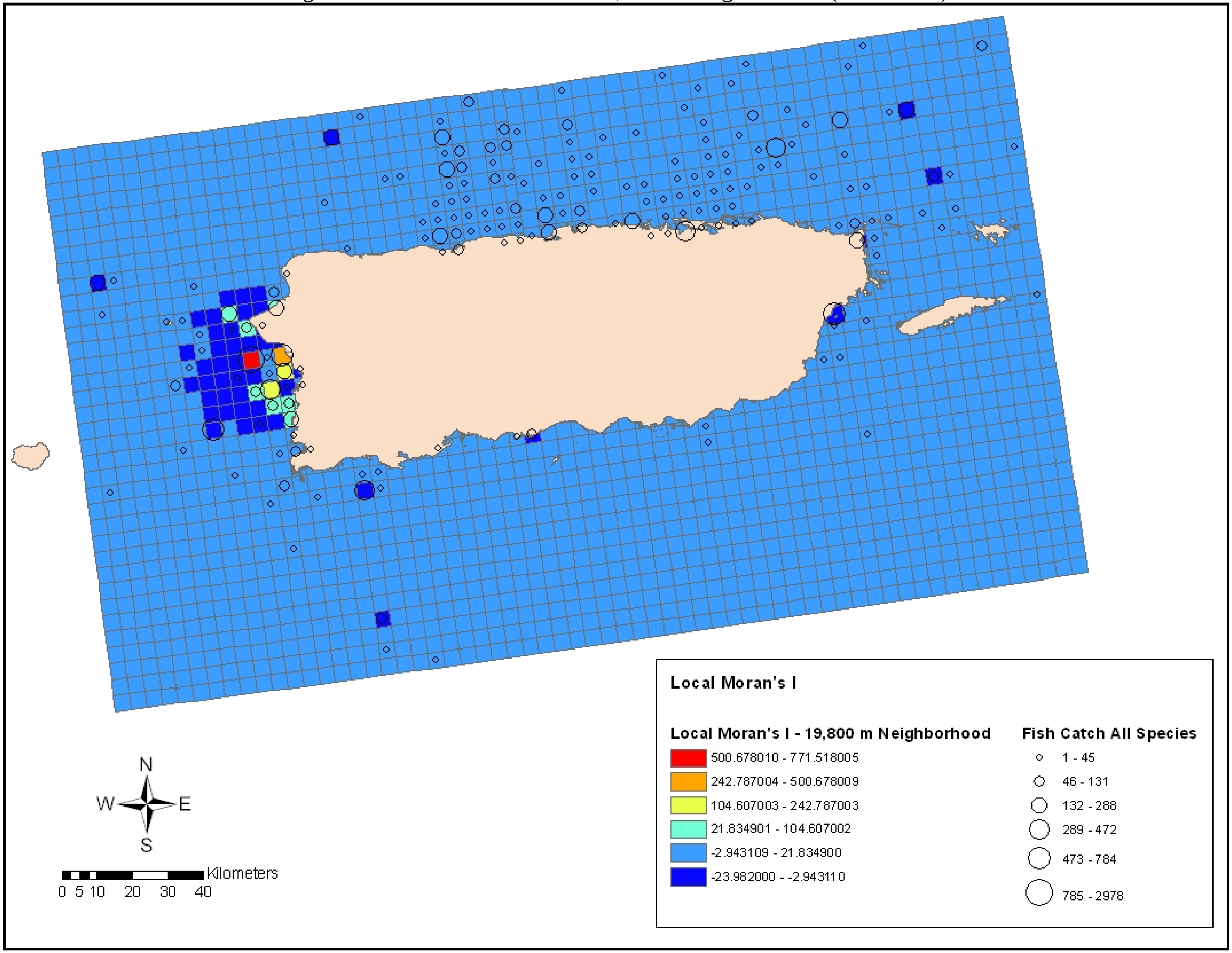


Figure 11: Local Moran's I with 4,900 m Neighborhood (1 centroid)
Symbolized By Standard Deviations from the Mean I Value

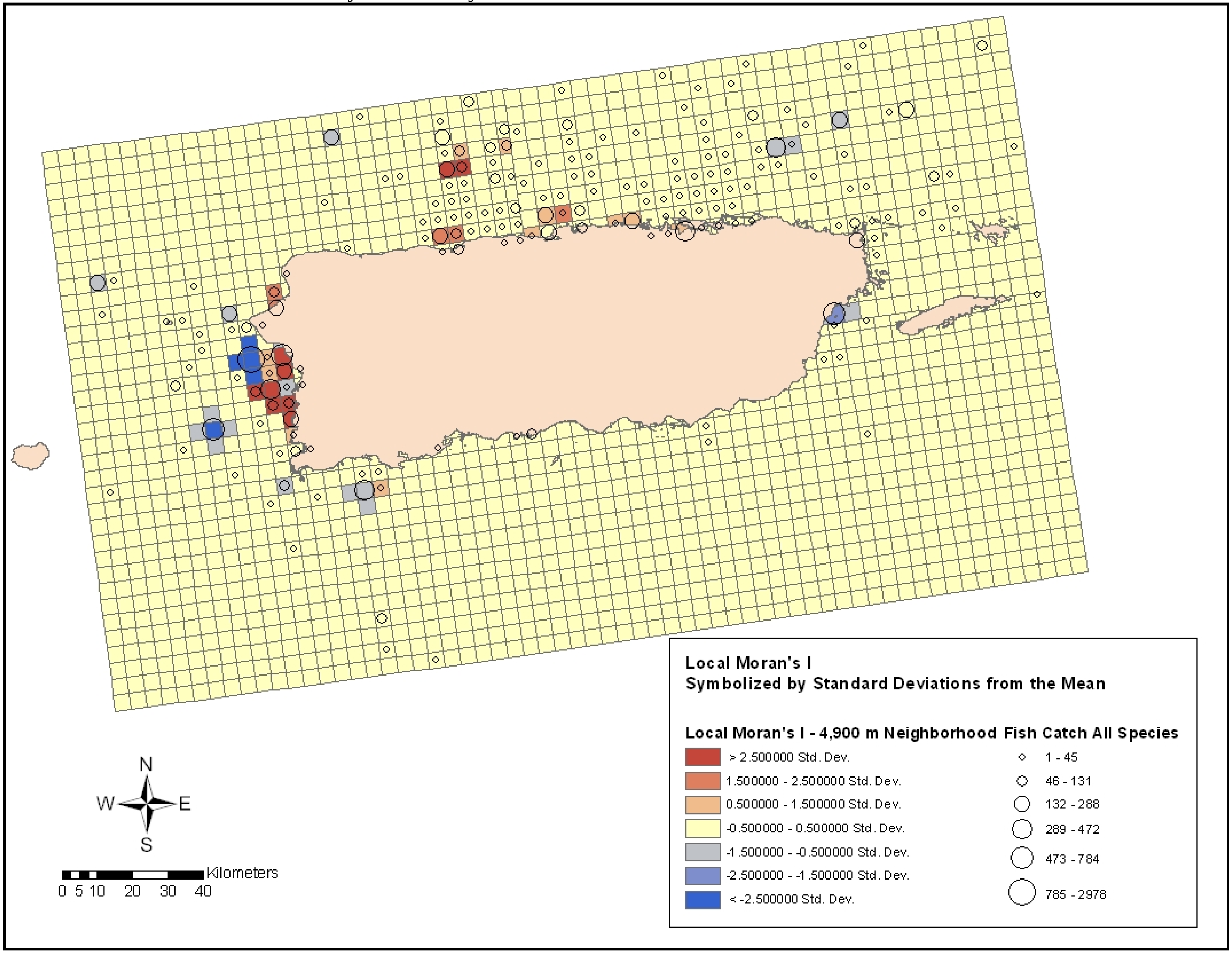


Figure 12: Local Moran's I with 9,800 m Neighborhood (2 centroids)
Symbolized By Standard Deviations from the Mean I Value

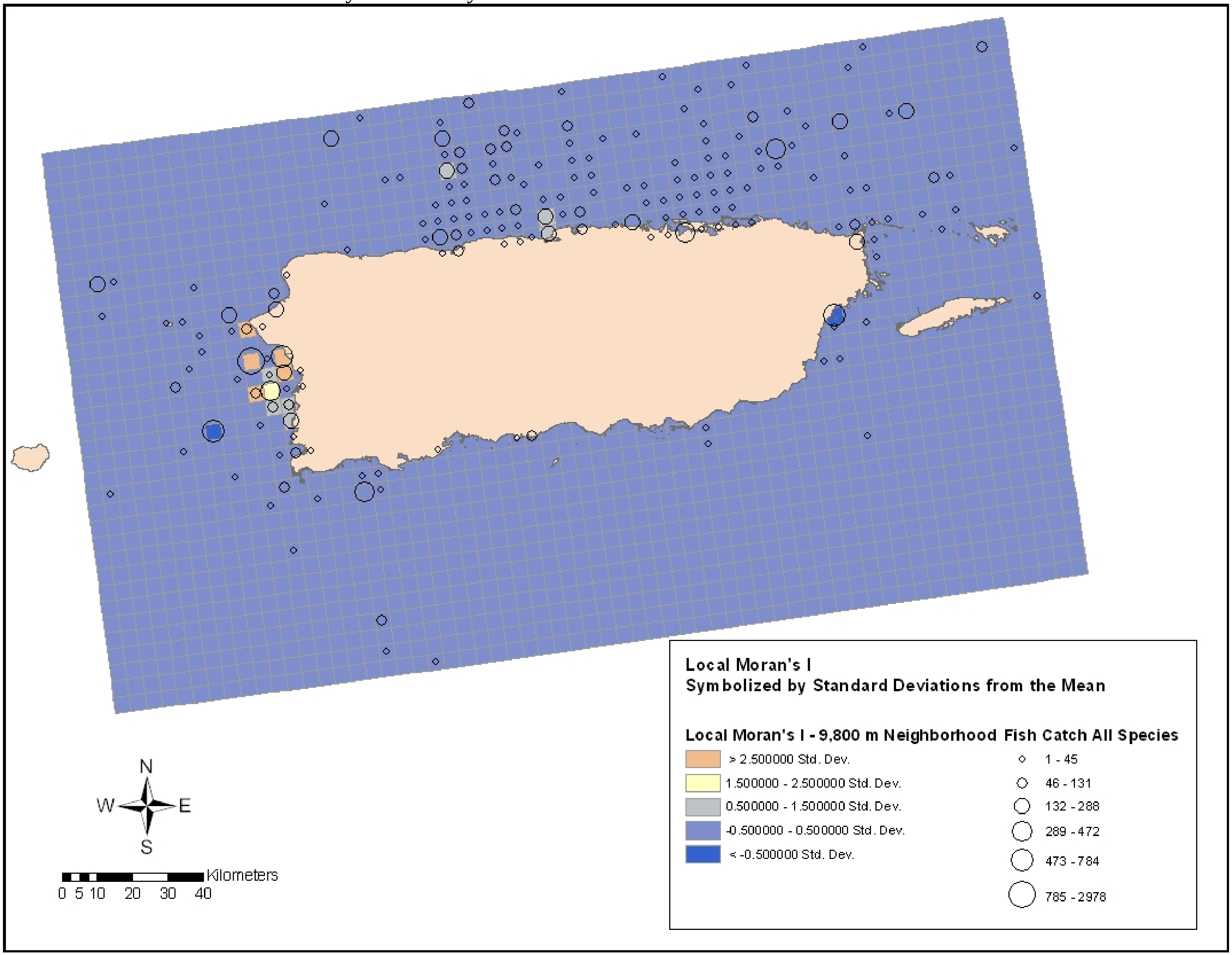


Figure 13: Local Moran's I with 14,700 m Neighborhood (3 centroids)
Symbolized By Standard Deviations from the Mean I Value

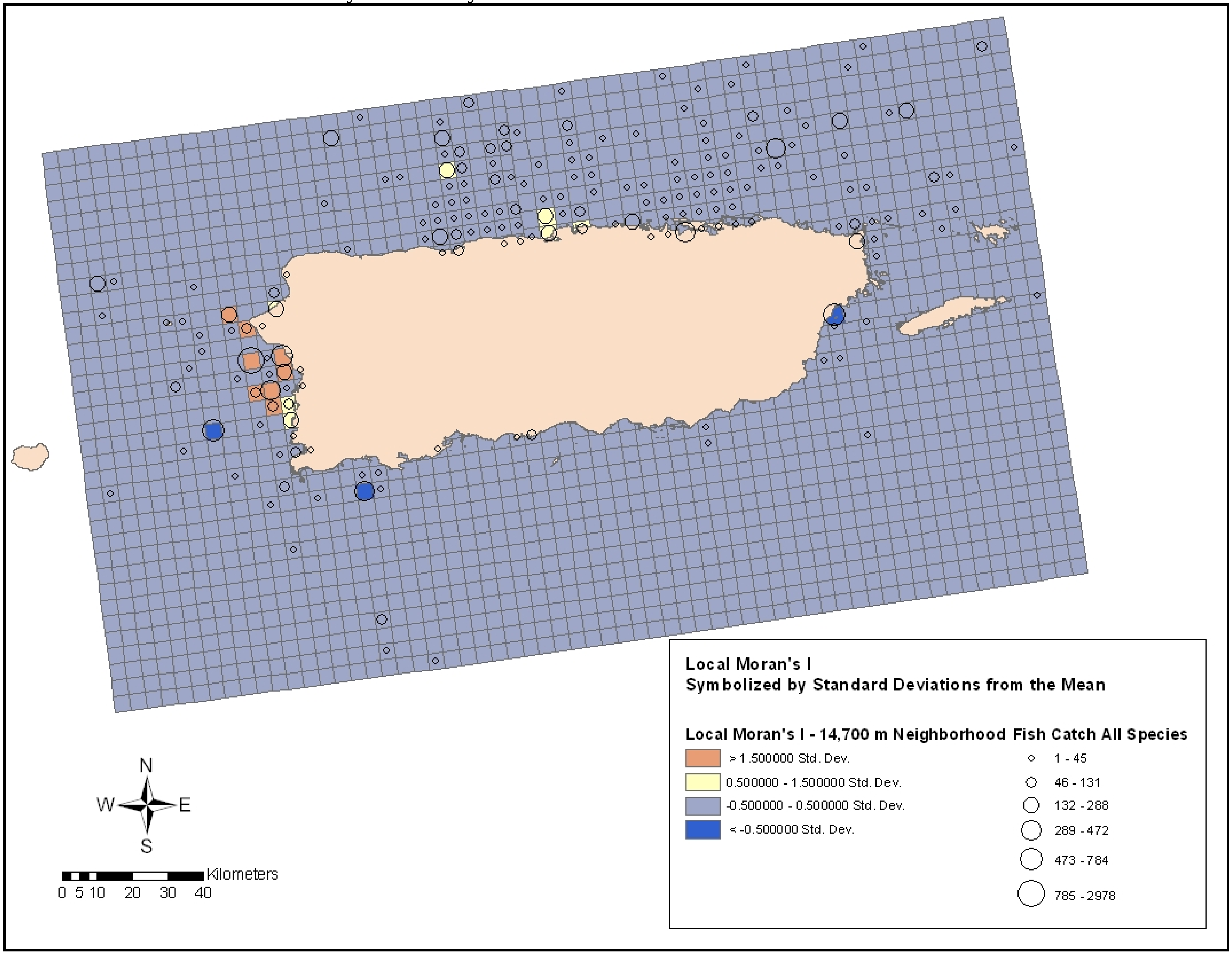


Figure 14: Local Moran's I with 19,800 m Neighborhood (4 centroids)
Symbolized By Standard Deviations from the Mean I Value

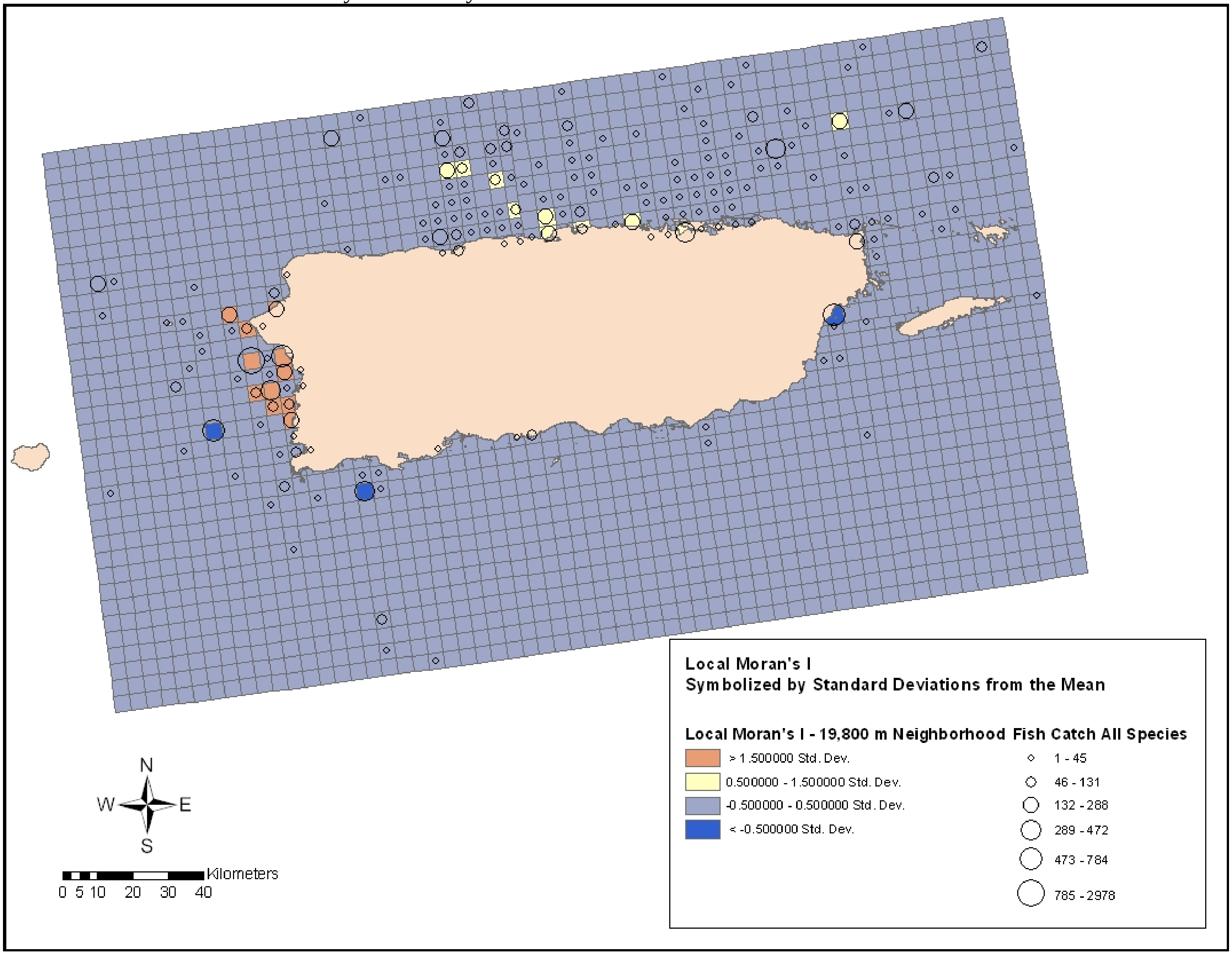


Figure 15: Local Moran's I Z-Score with 4,900m Neighborhood (1 Centroid)

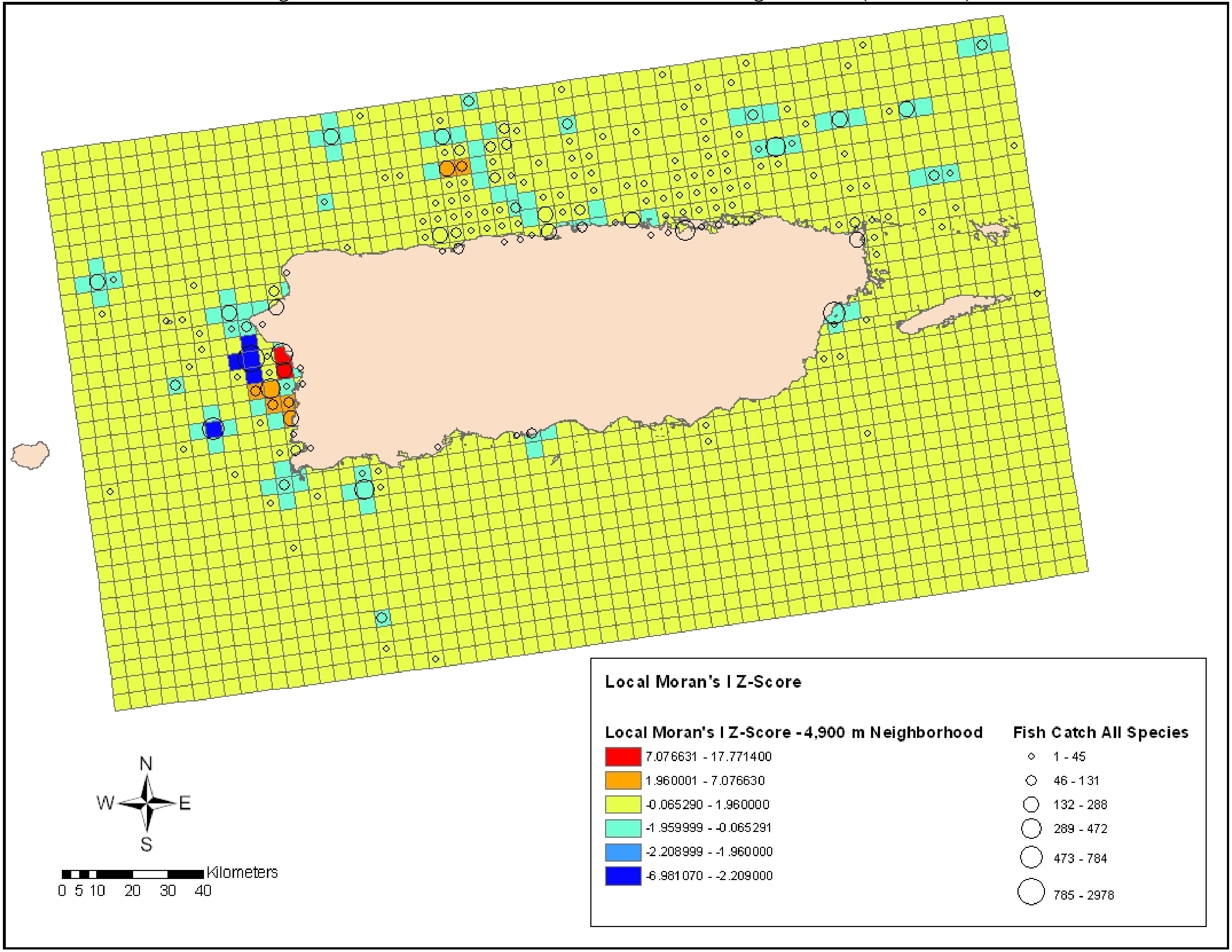


Figure 16: Local Moran's I Z-Score with 9,800m Neighborhood (2 Centroids)

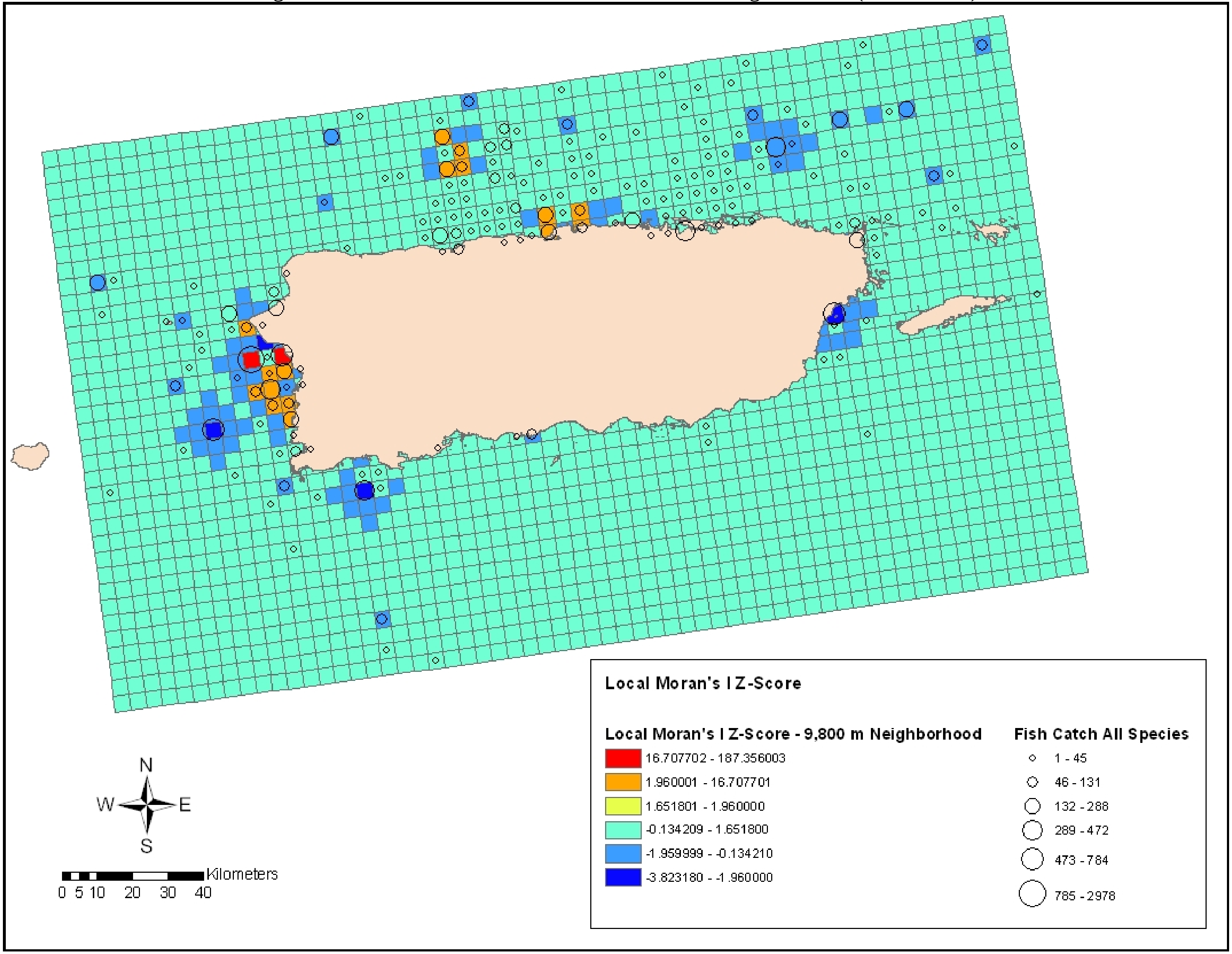


Figure 17: Local Moran's I Z-Score with 14,700m Neighborhood (3 Centroids)

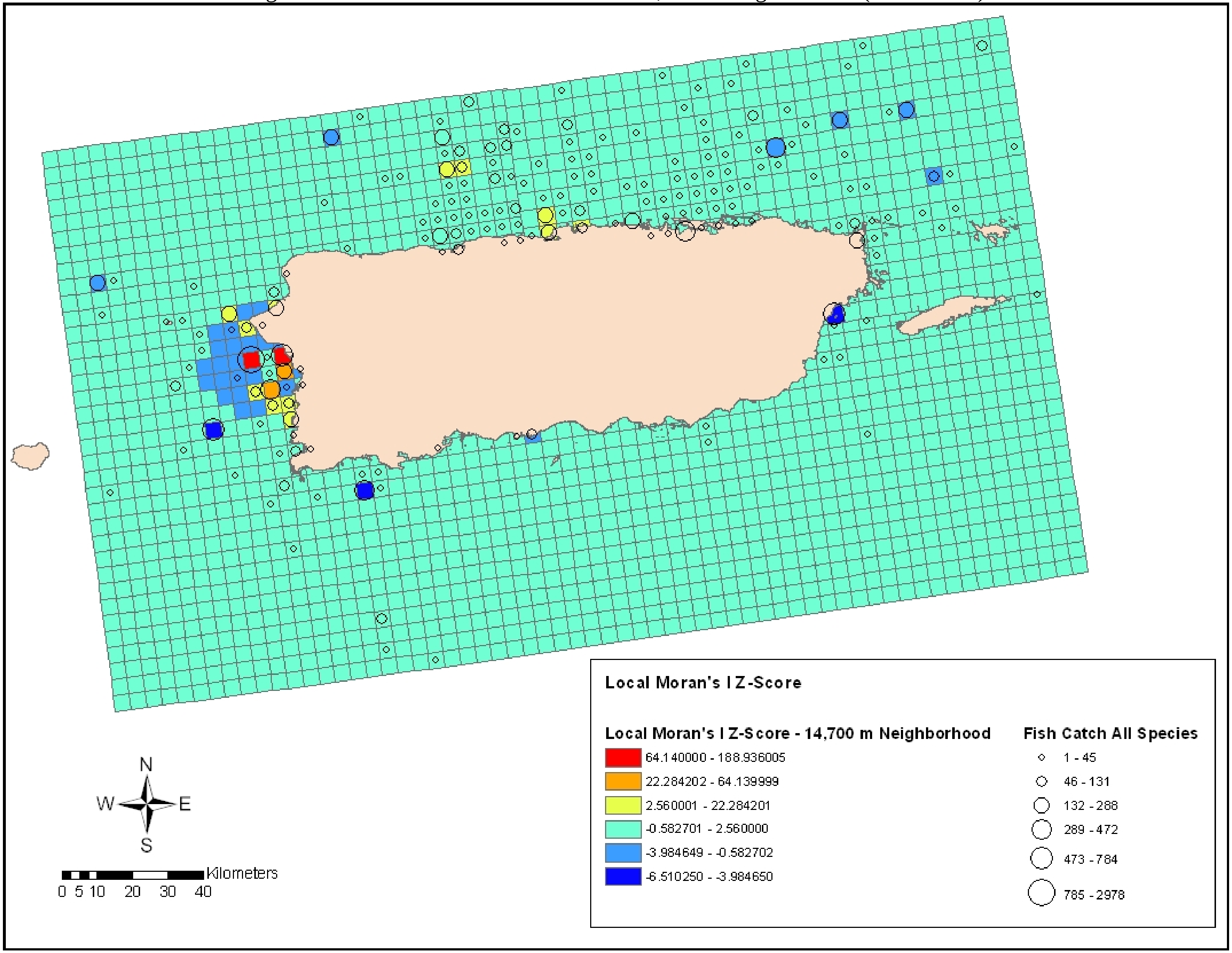


Figure 18: Local Moran's I Z-Score with 19,800m Neighborhood (4 Centroids)

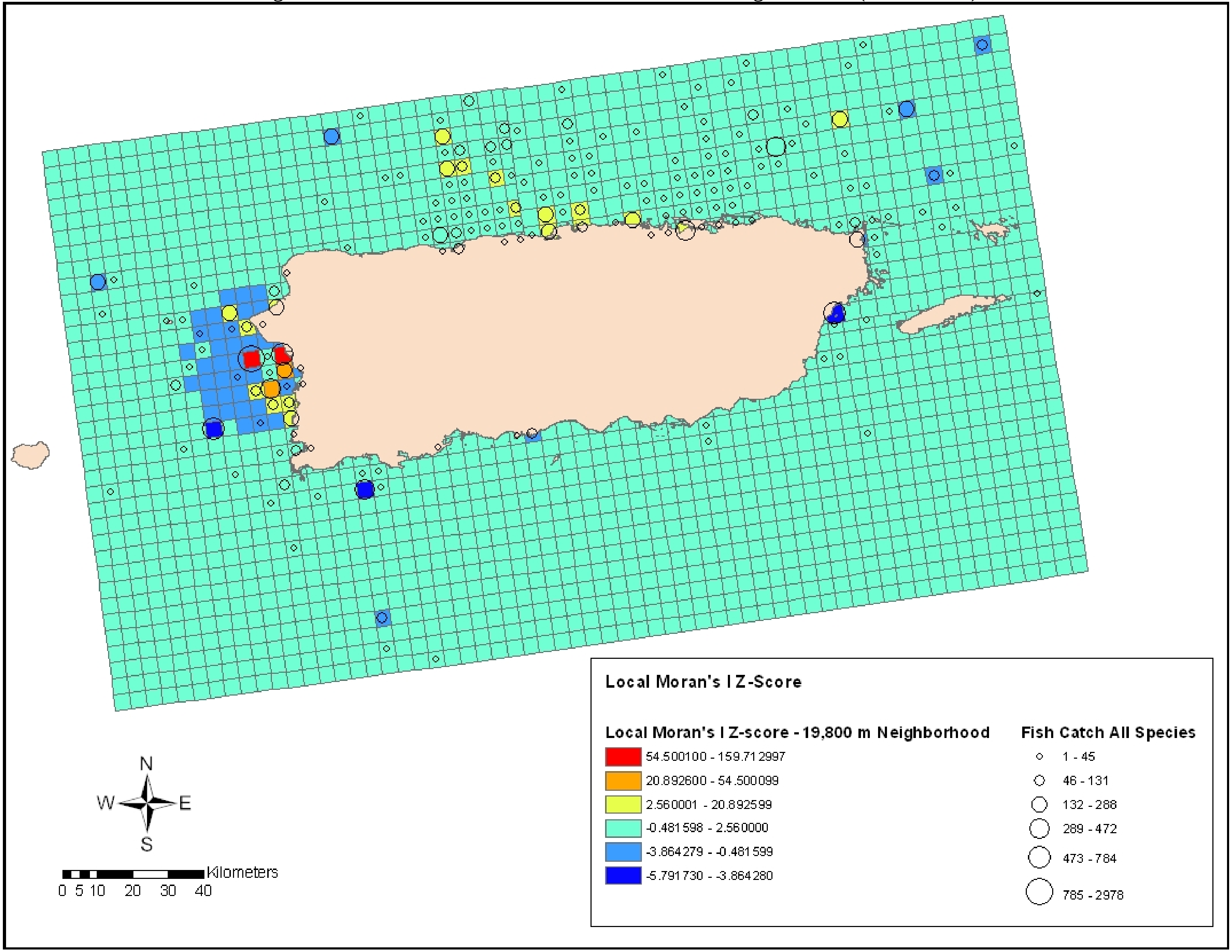


Figure 19: Gi*Z-Score with 4,900 m Neighborhood (1 centroid)

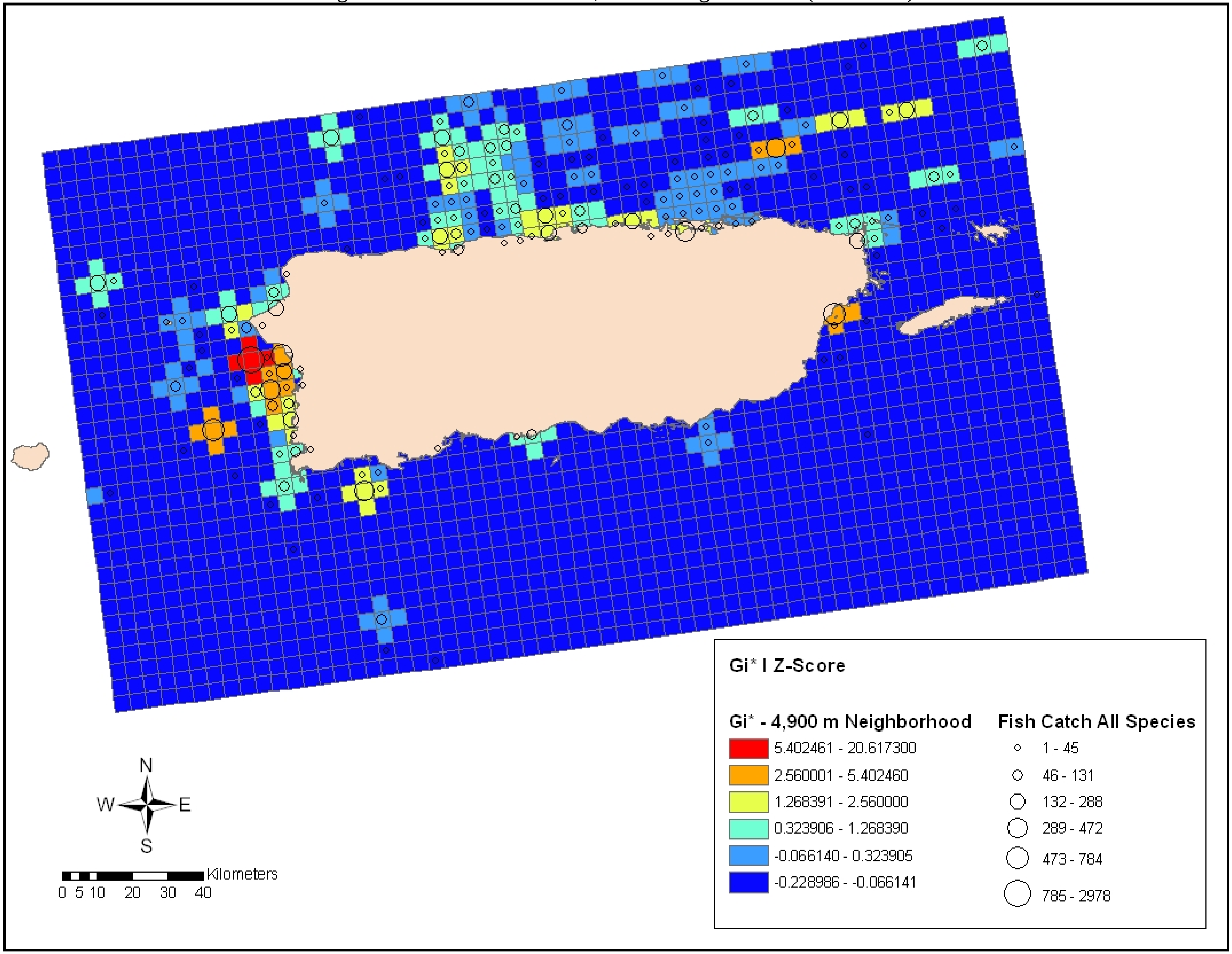


Figure 20: Gi* Z-Score with 9,800 m Neighborhood (2 centroids)

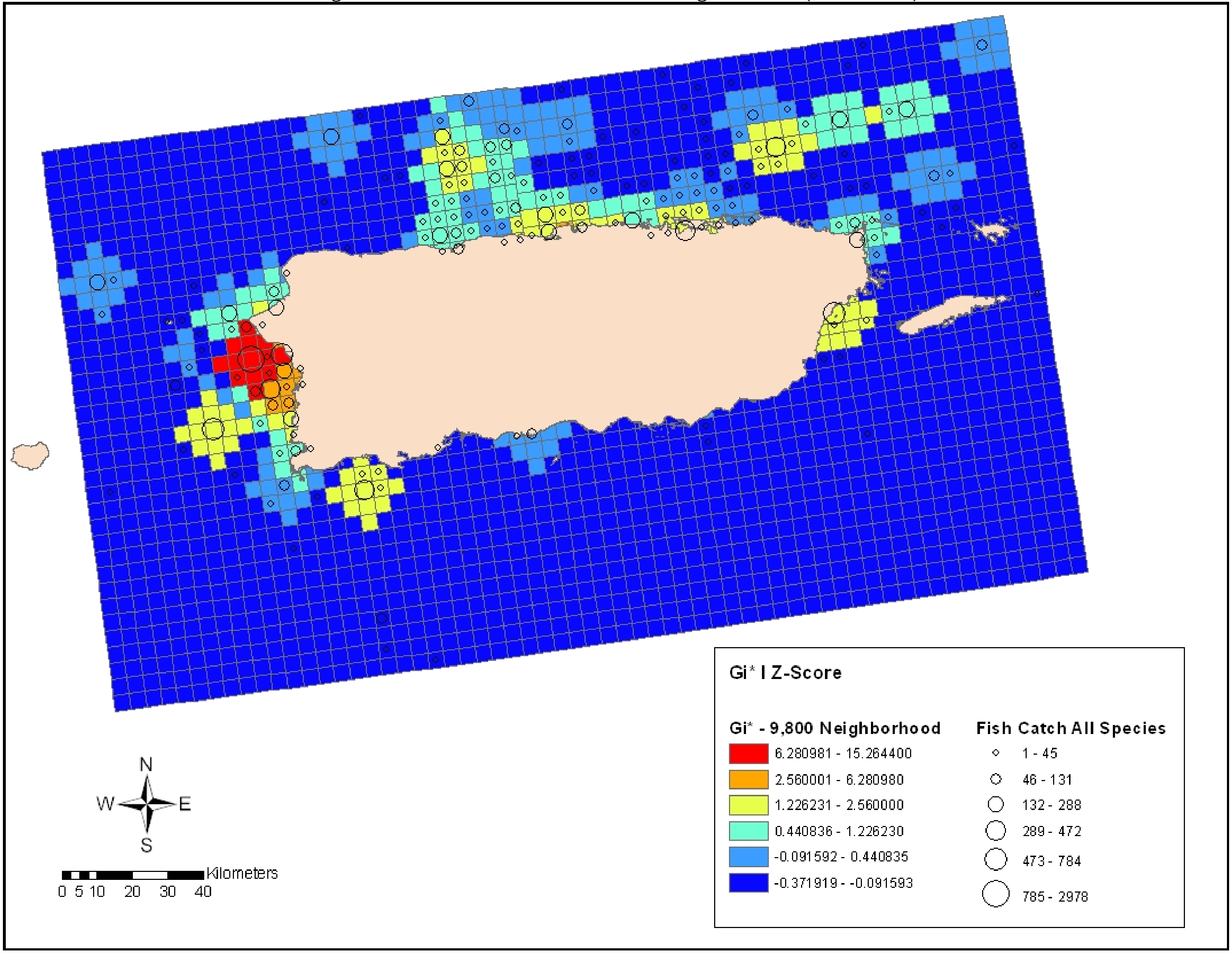


Figure 21: Gi* Z-Score with 14,700 m Neighborhood (3 centroids)

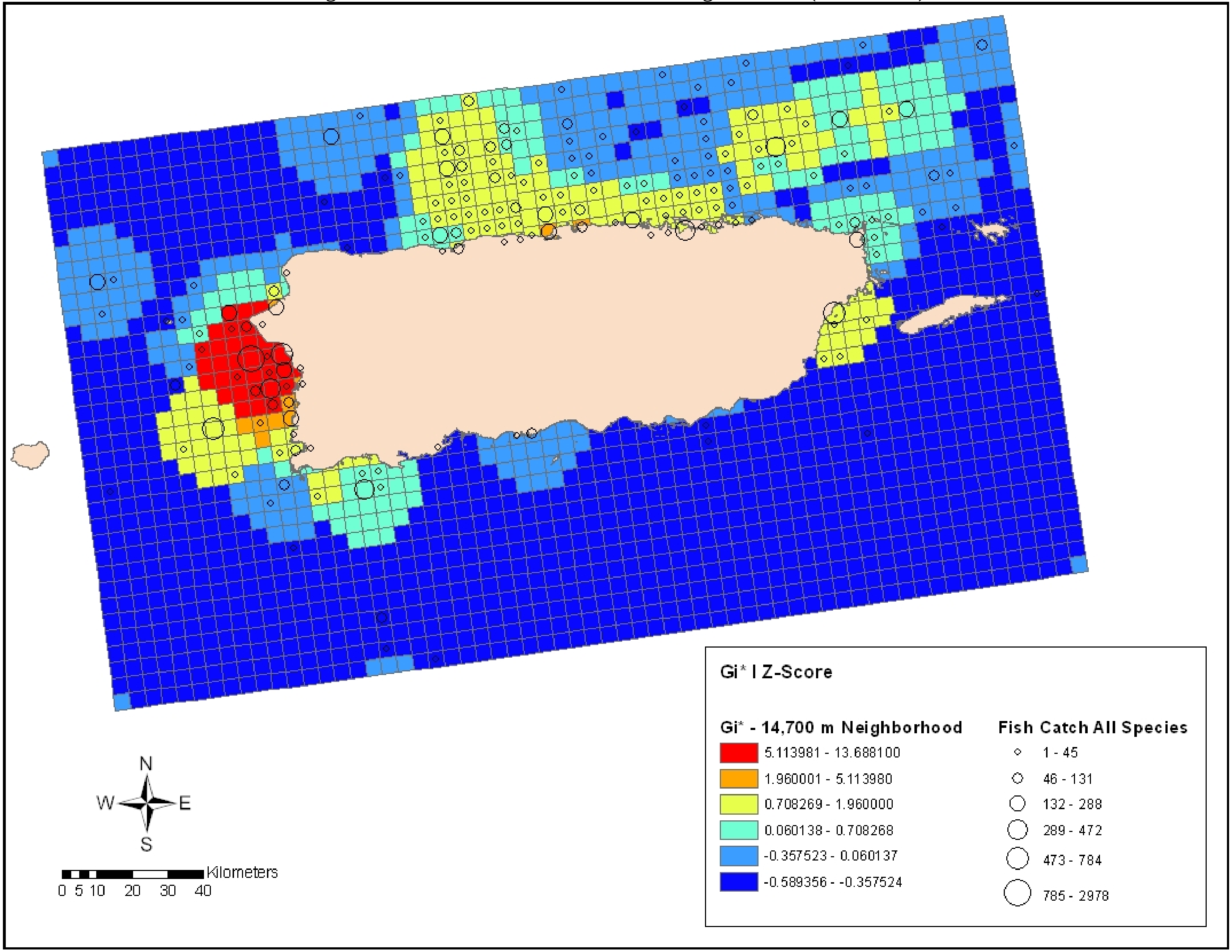


Figure 22: Gi* Z-Score with 19,800 m Neighborhood (4 centroids)

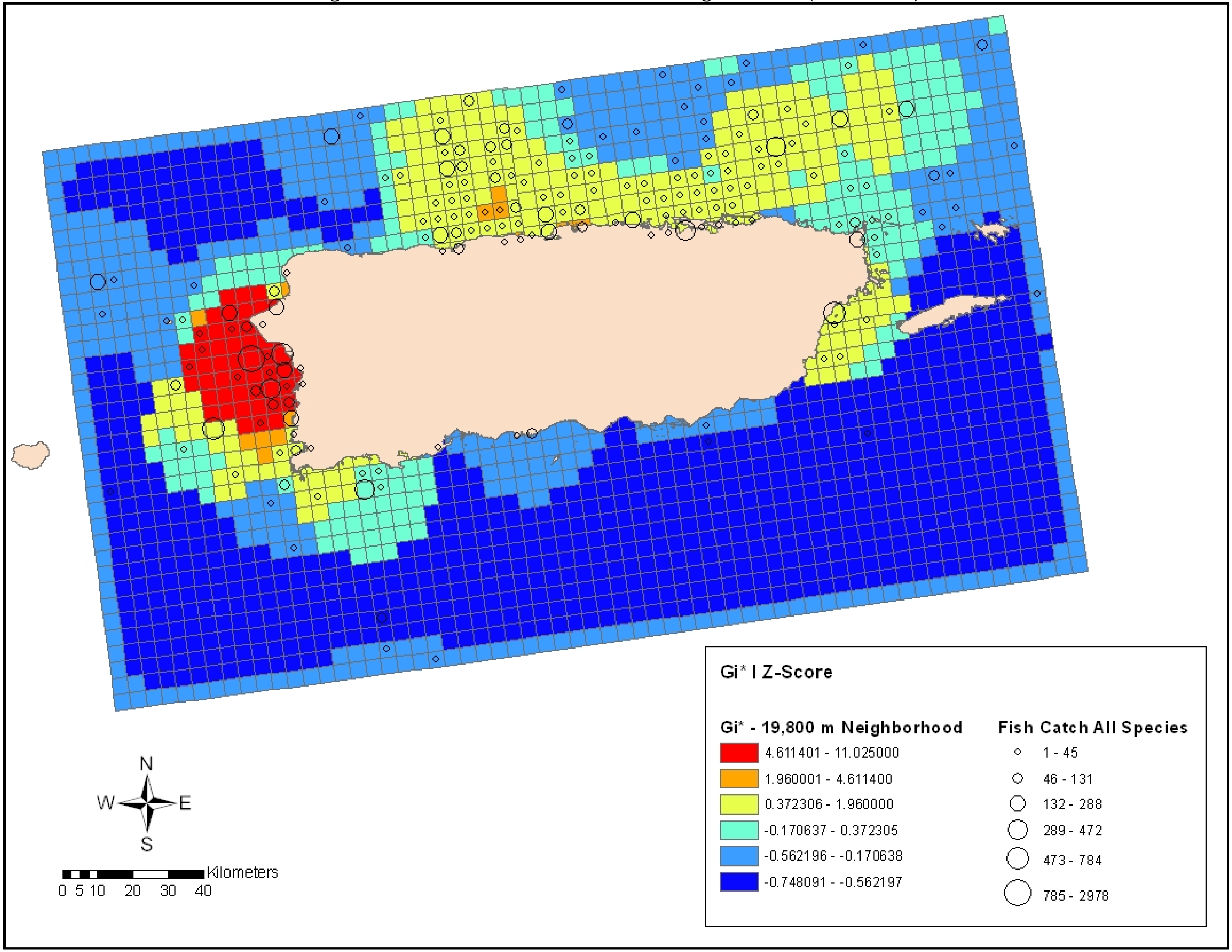


Figure 23: Dolphin Local Moran's I with 4,900 m Neighborhood (1 centroid)

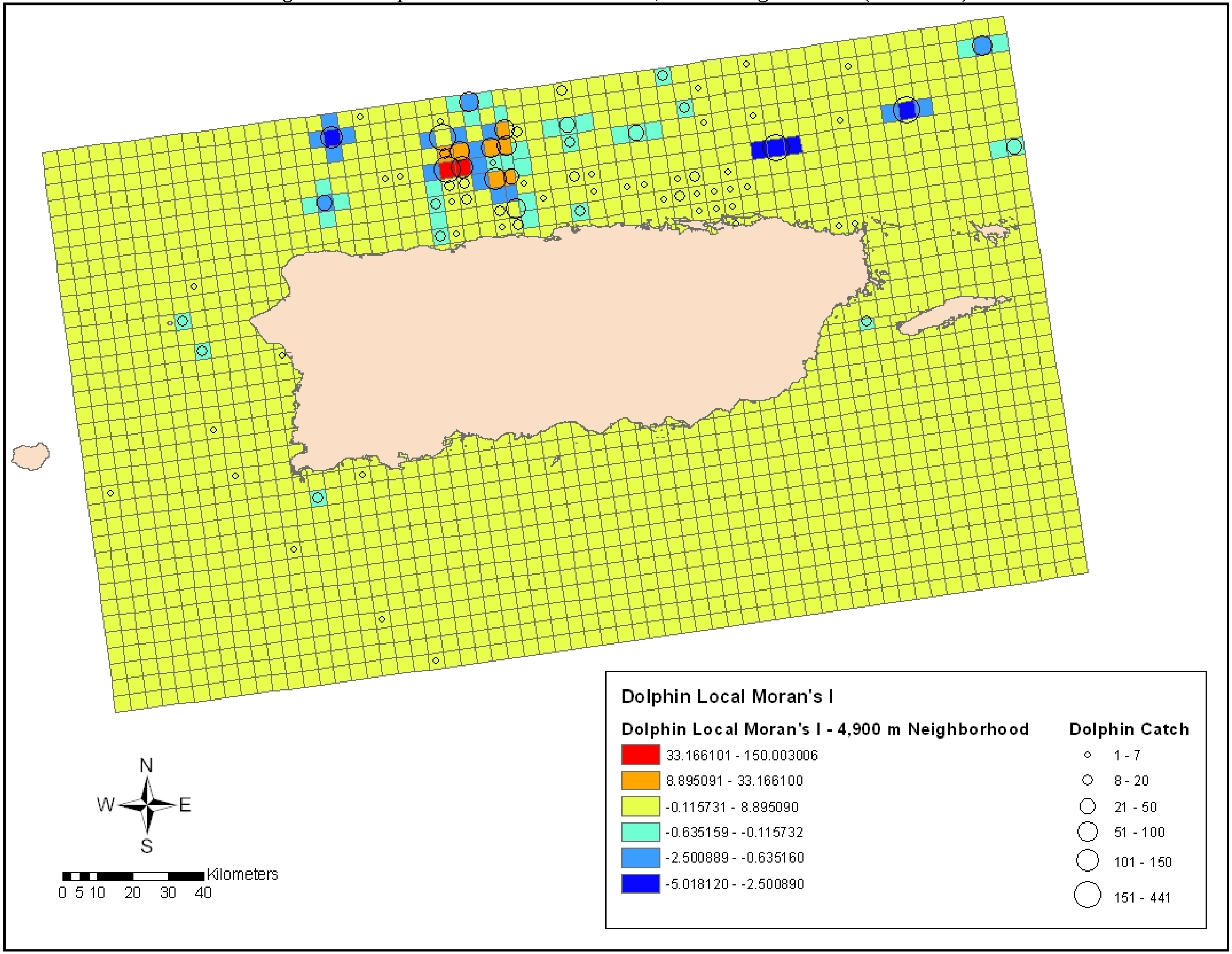


Figure 24: Dolphin Local Moran's I with 9,800 m Neighborhood (2 centroids)

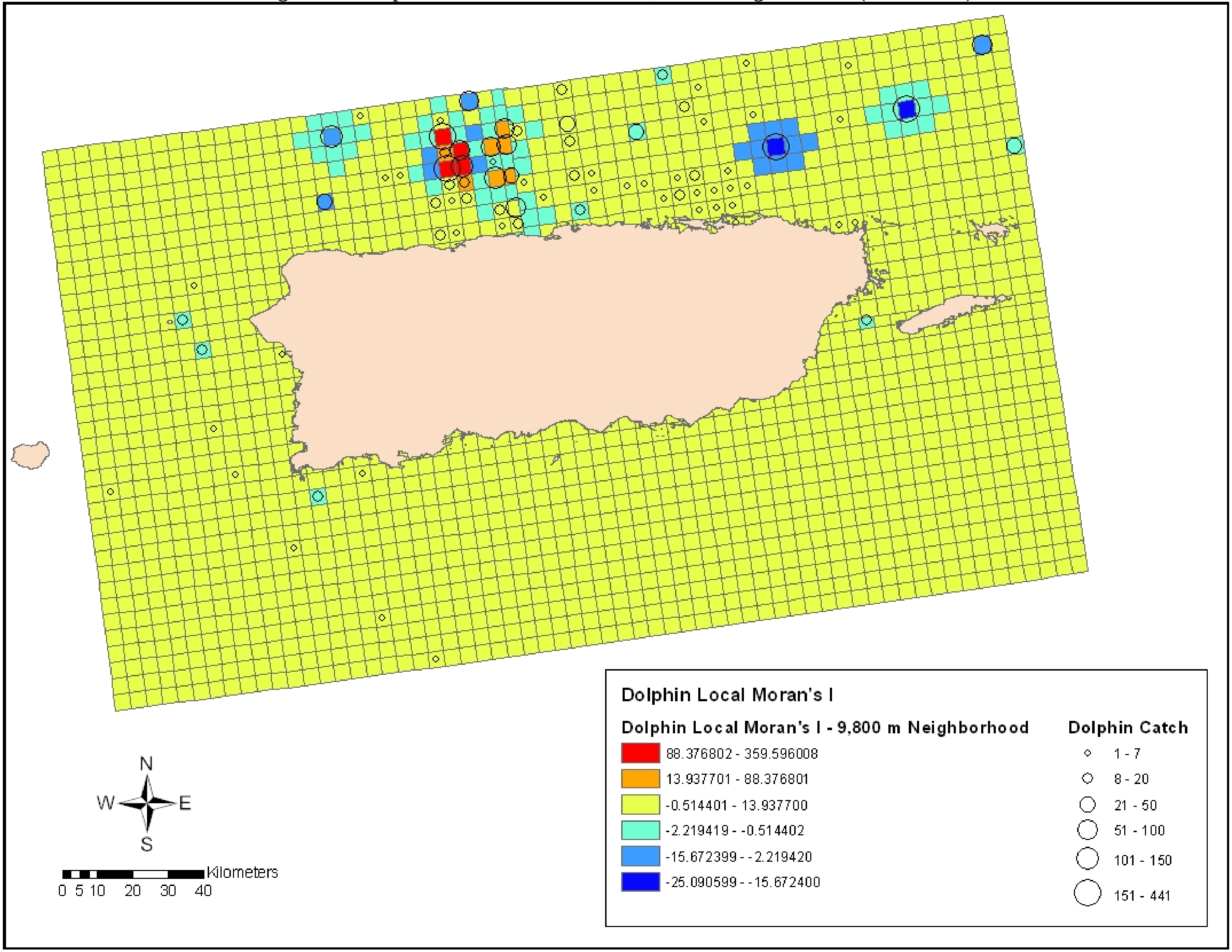


Figure 25: Dolphin Local Moran's I with 14,700 m Neighborhood (3 centroids)

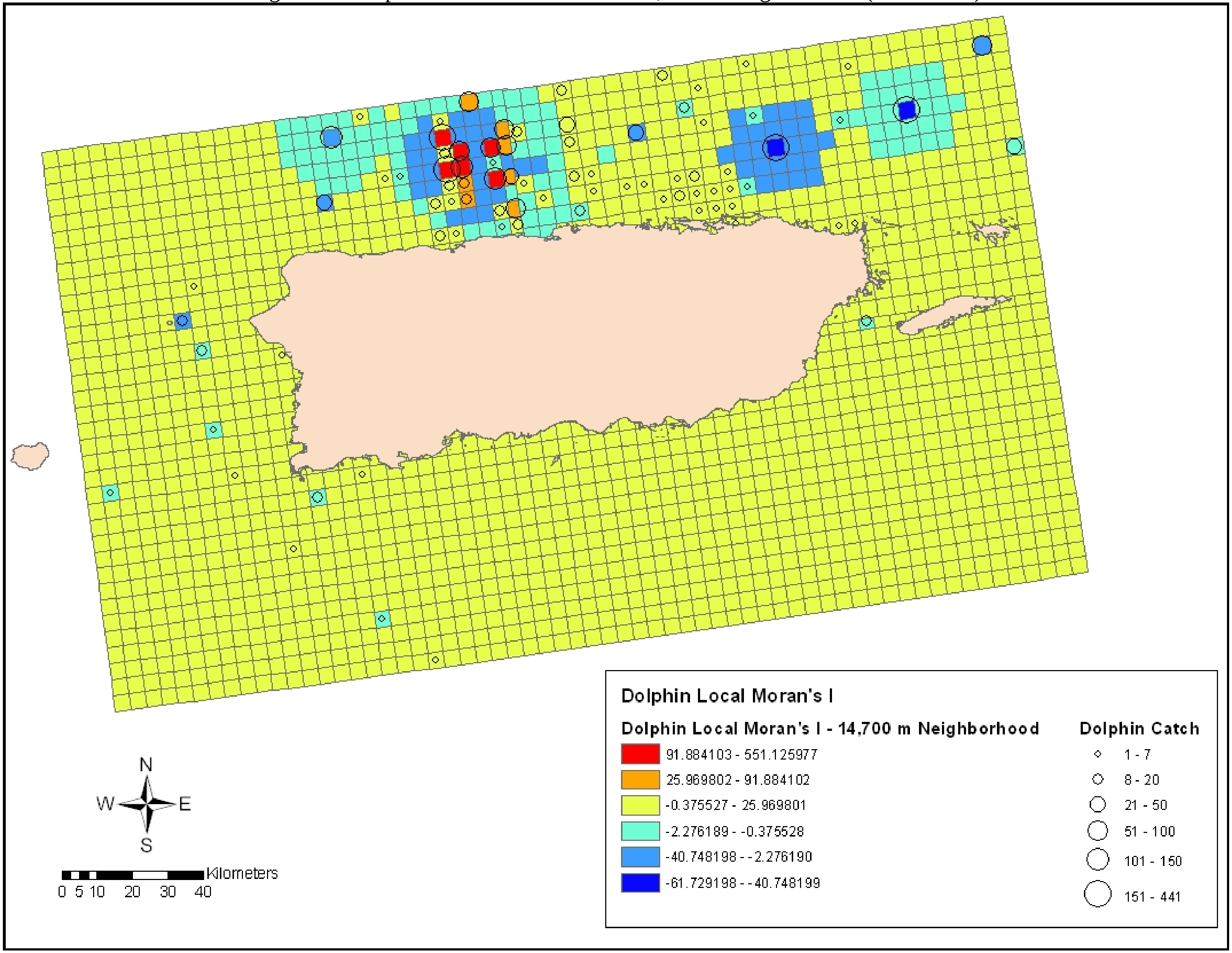


Figure 26: Dolphin Local Moran's I with 19,800 m Neighborhood (4 centroids)

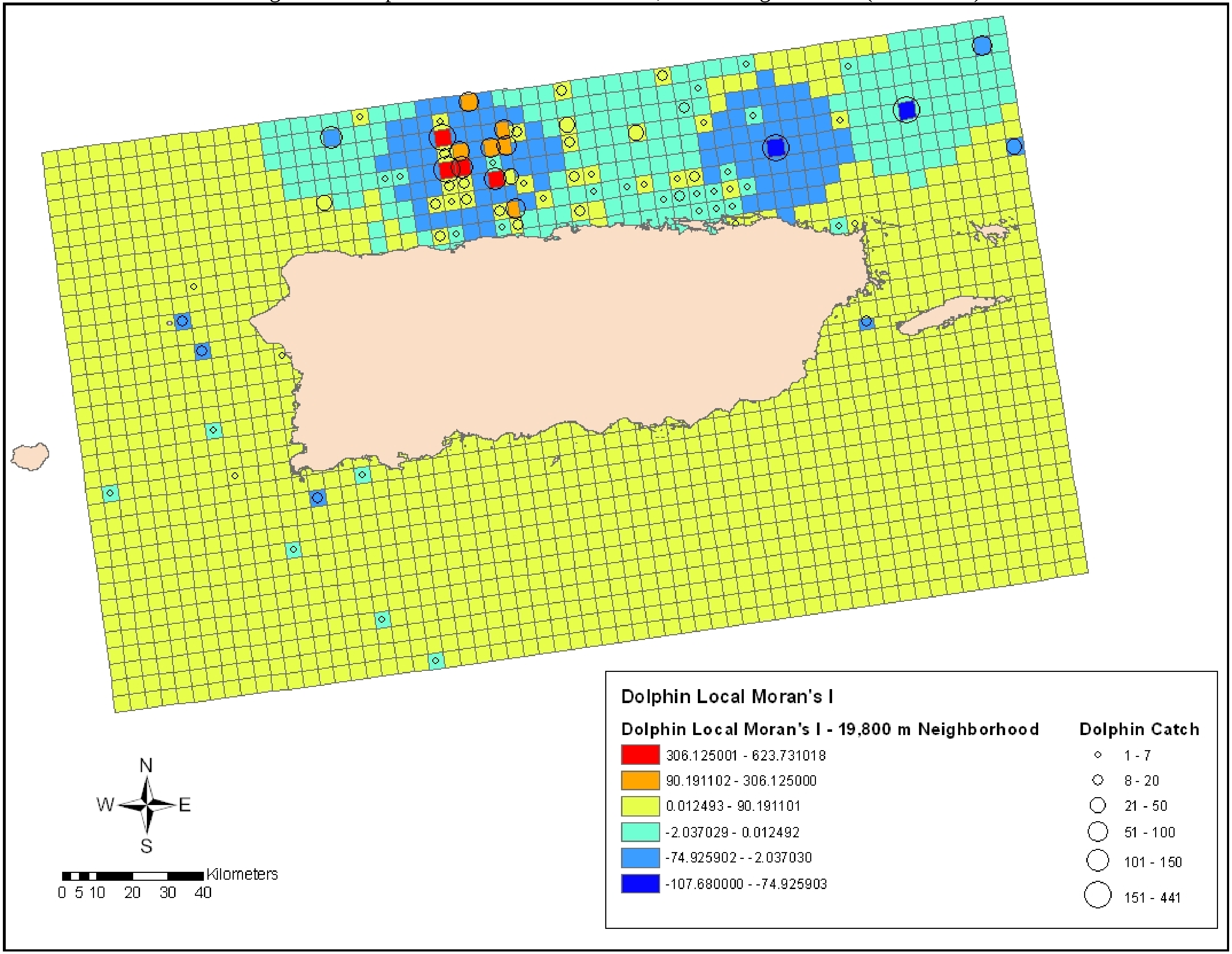


Figure 27: Dolphin Local Moran's I Z-score with 4,900 m Neighborhood (1 centroid)

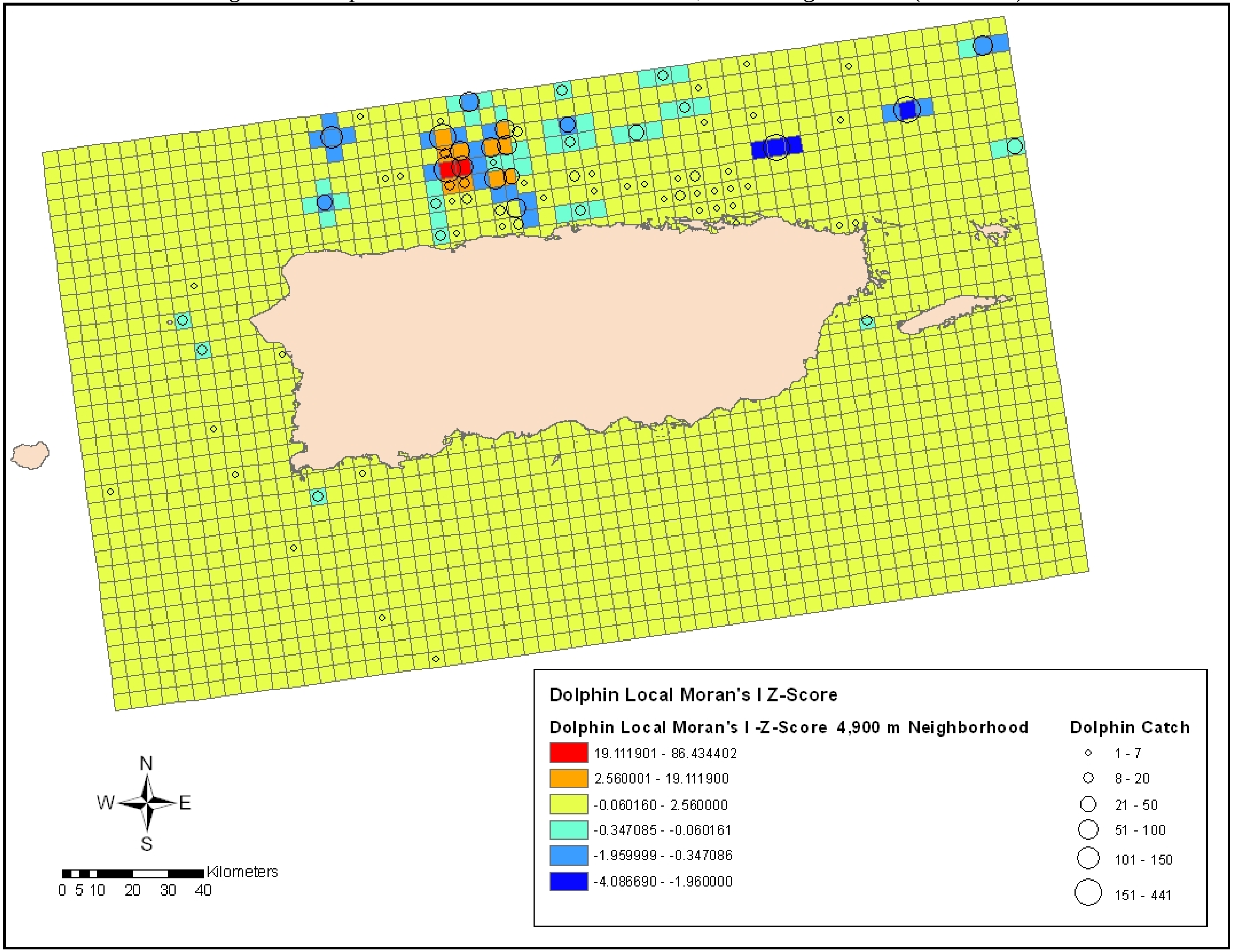


Figure 28: Dolphin Local Moran's I Z-score with 9,800 m Neighborhood (2 centroids)

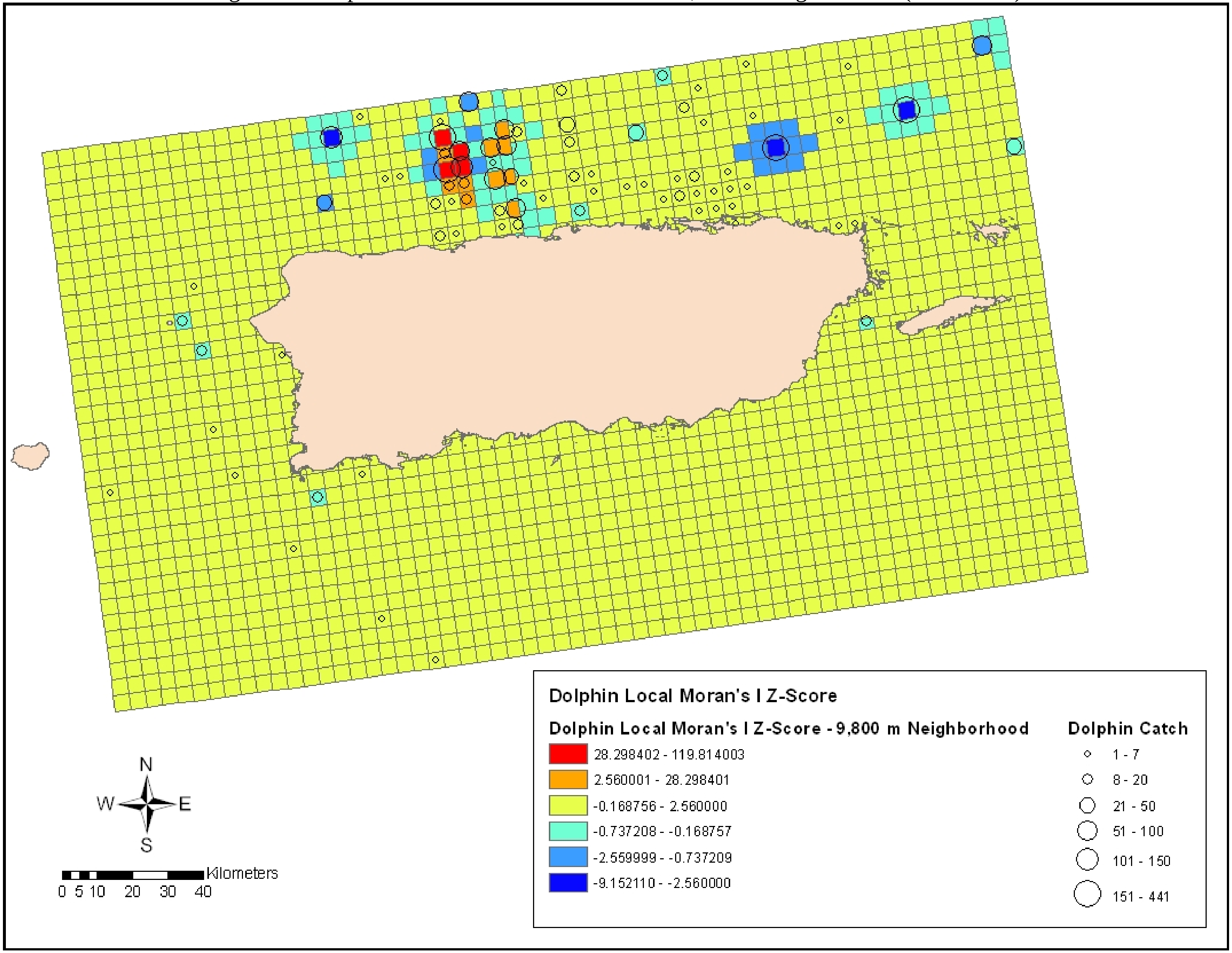


Figure 29: Dolphin Local Moran's I Z-score with 14,700 m Neighborhood (3 centroids)

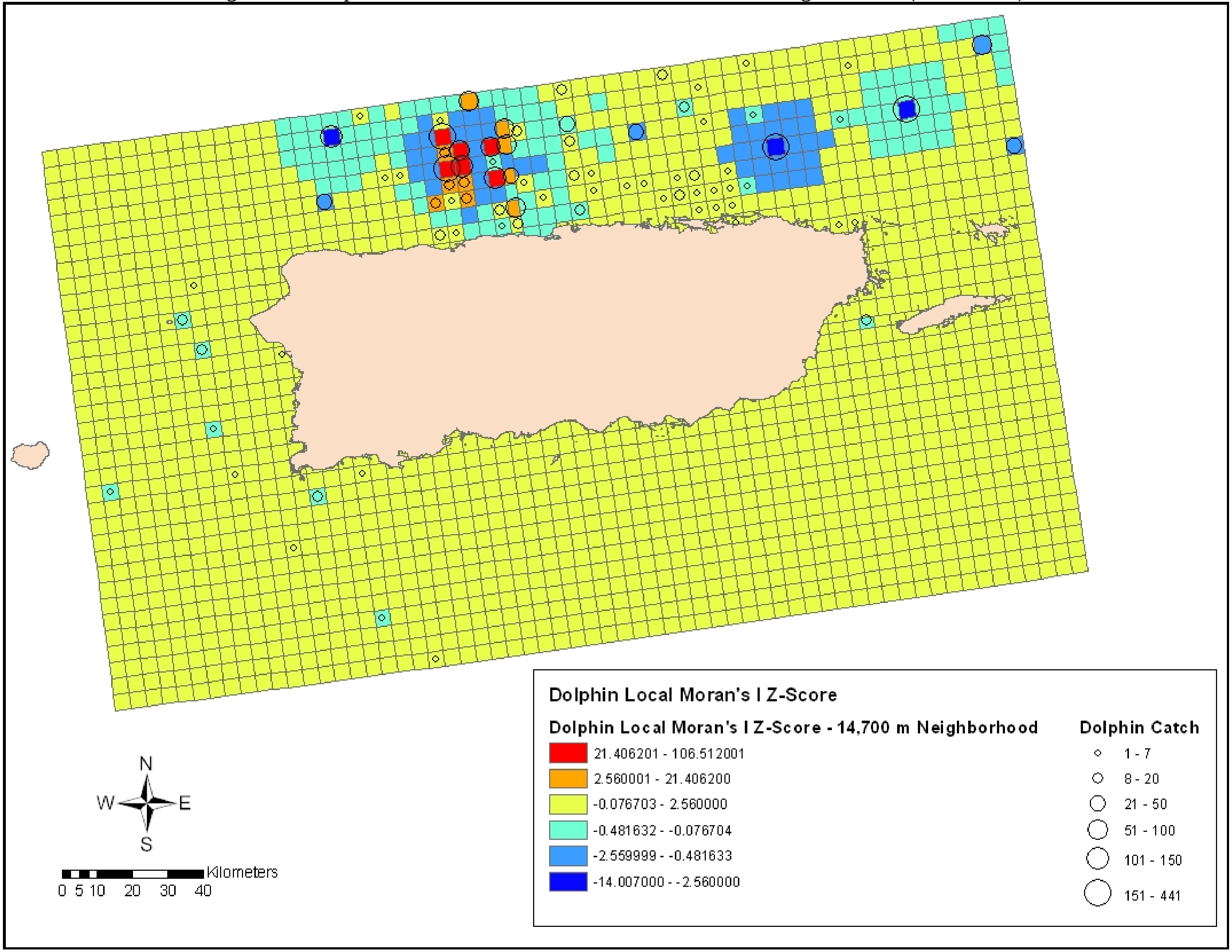


Figure 30: Dolphin Local Moran's I Z-score with 19,800 m Neighborhood (4 centroids)

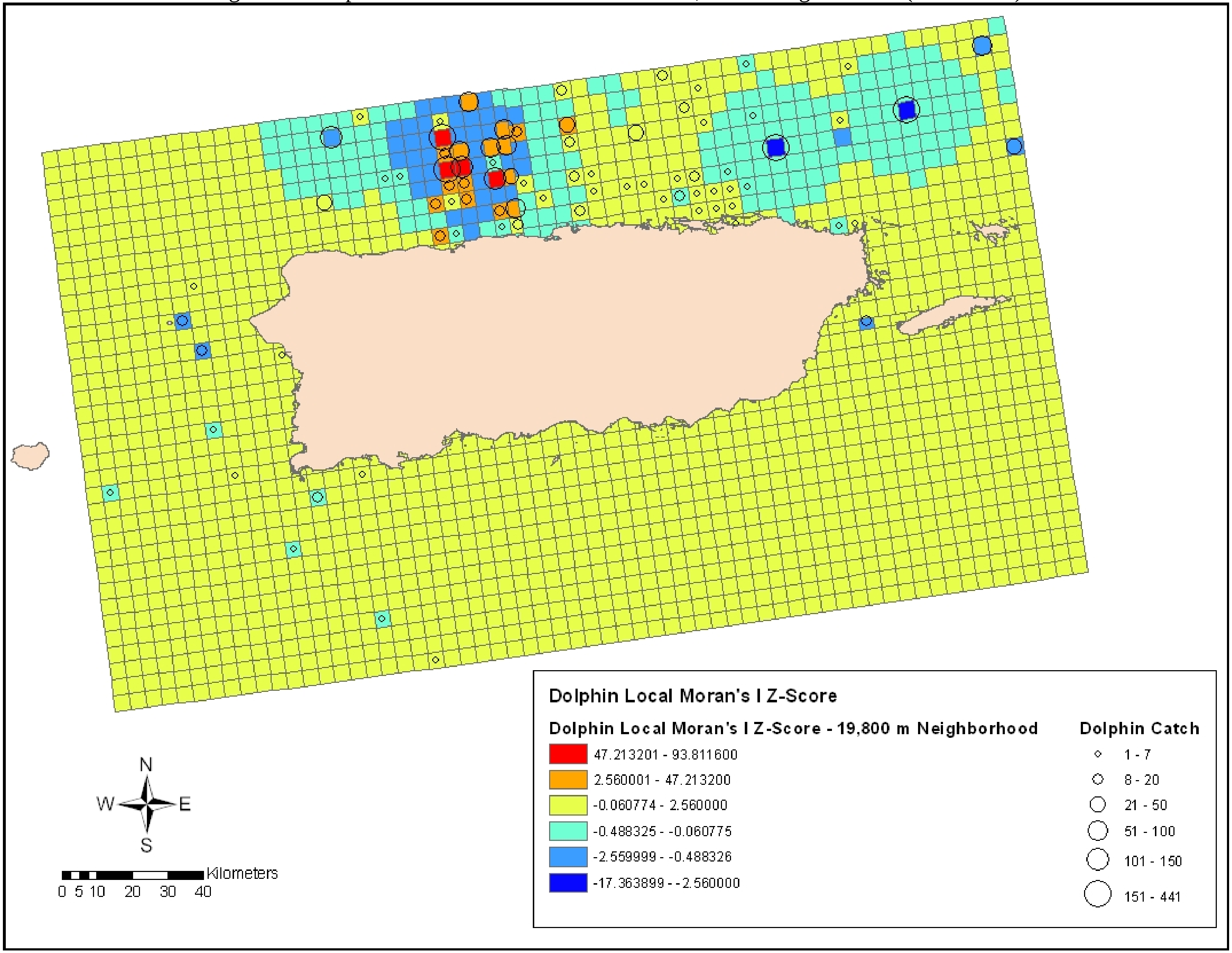


Figure 31: Dolphin Gi* Z-score with 4,900 m Neighborhood (1 centroid)

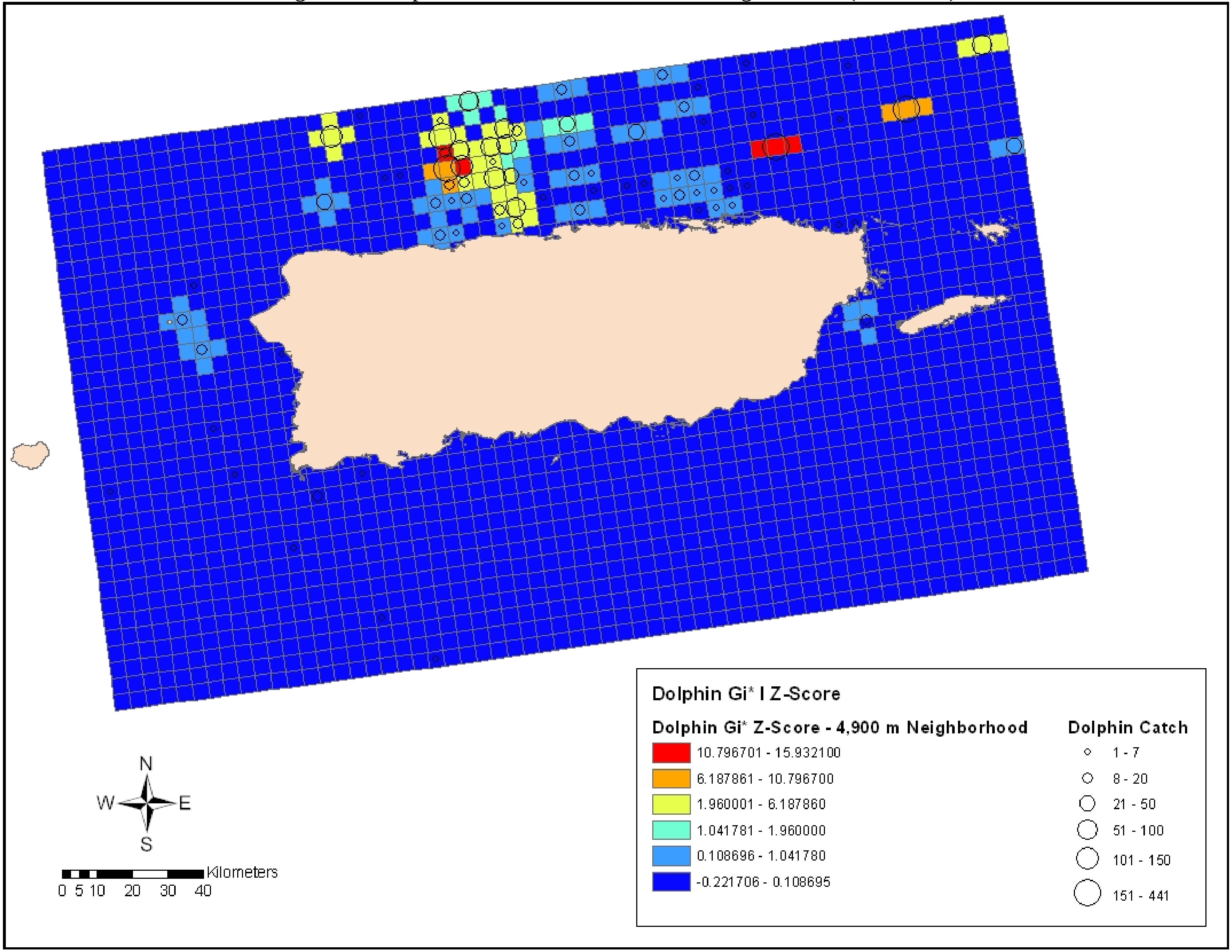


Figure 32: Dolphin Gi*I Z-score with 9,800 m Neighborhood (2 centroids)

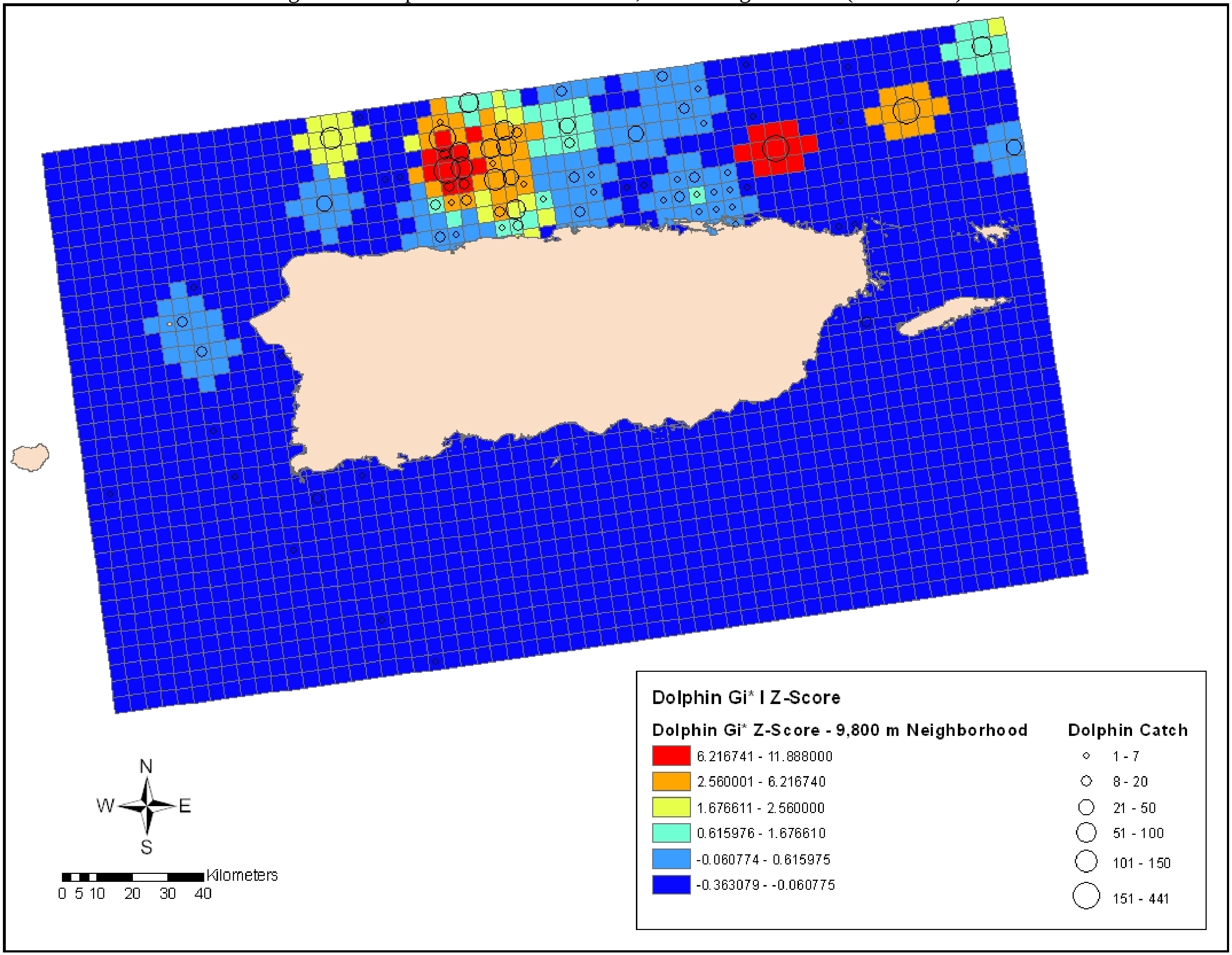


Figure 33: Dolphin Gi* Z-score with 14,700 m Neighborhood (3 centroids)

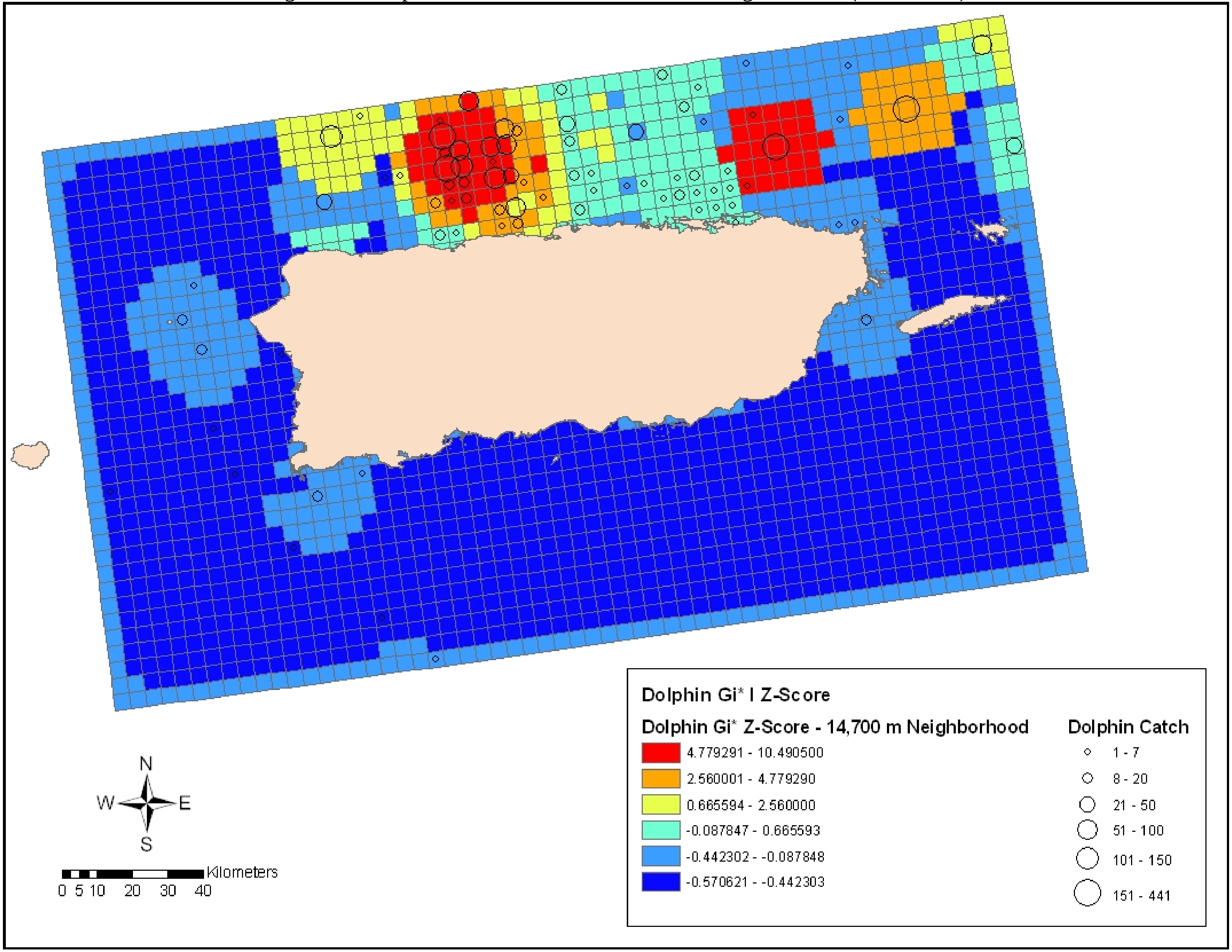
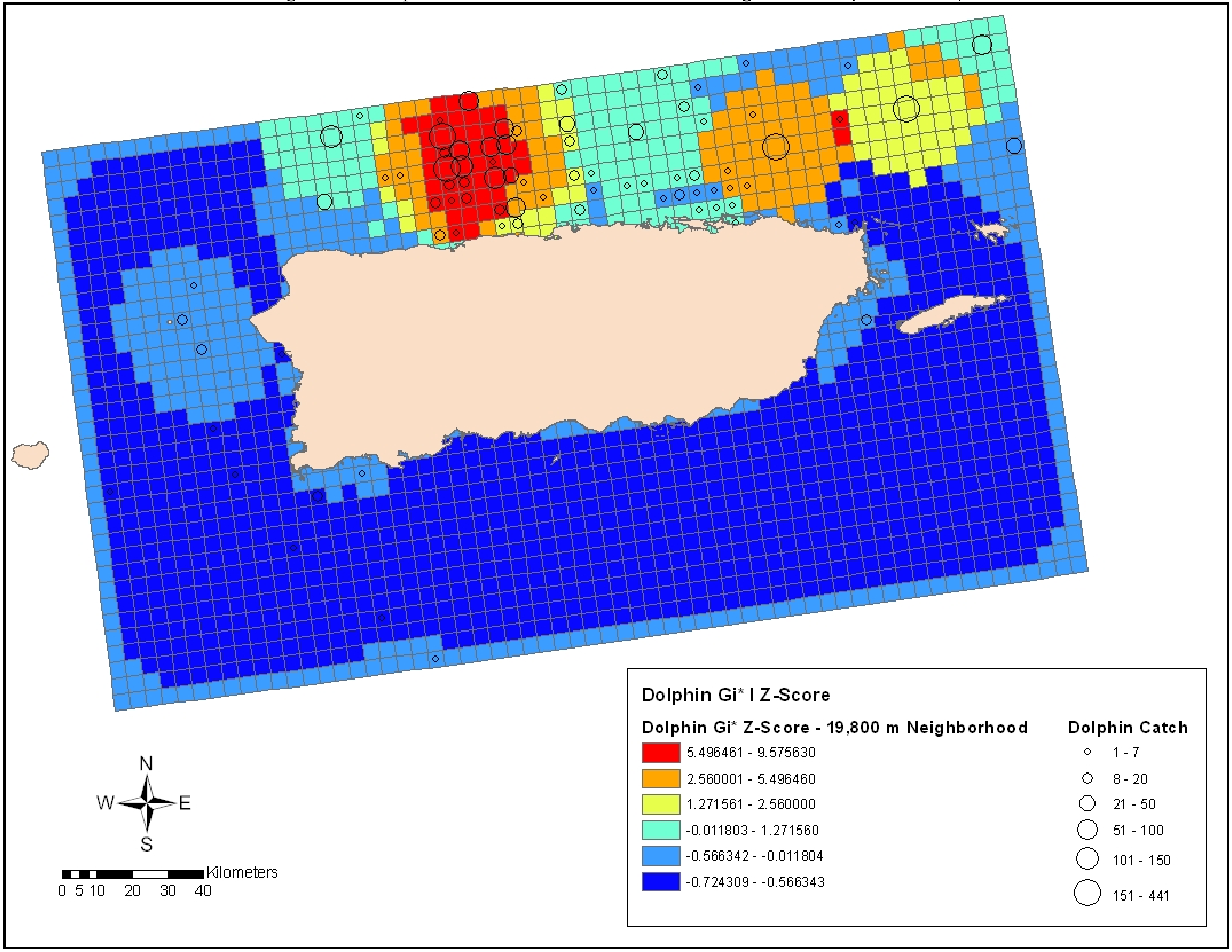


Figure 34: Dolphin Gi* Z-score with 19,800 m Neighborhood (4 centroids)



Metadata

The following section details the data used and created during the GIS analysis of this project.

Table 7: Description of primary GIS tables

Table Name	*Description*
ETgrid	Data from the 2.5 minute grid shapefile
nodcodes_cat	12-digit NODC species codes used in the MRFSS
i1g_pr	Type 1 MRFSS data: trip and angler characteristics
i2g_pr	Type 2 MRFSS data: unobserved, reported catch
i3g_pr	Type 3 MRFSS data; inspected catch
i5g_pr	Type 5 data: sociological information, including area fished

Table 8: Description of summary and joined tables created for GIS analysis

Table Name	*Description*
table2_reported	Data from table i2g_pr joined to 12-digit NODC species code.
table2_grid	Data from table2_reported joined to i5g_pr table by the ID_CODE field. Resulting table contains GRIDKECH field for spatial analysis.
table2_sum	Data from table2_grid summarized by GRIDKECH and NUM_FISH fields.
table3_reported	Data from table i3g_pr joined to 12-digit NODC species code.
table3_grid	Data from table3_reported joined to i5g_pr table by the ID_CODE field. Resulting table contains GRIDKECH field for spatial analysis.
table3_sum	Data from table3_grid summarized by GRIDKECH and FSHINSP fields.

The data tables listed below were created for dolphin, lane snapper, snook, yellowtail snapper, and wahoo species. Species* is used in place of the specific fish species to simplify the summary.

Table 9: Species data tables

Table Name	*Description*
species*_table2	All species* data based on selection of unique NODC code from table2_reported .
species*_table2_grid	Data from species*_table2 joined to i5g_pr table by the ID_CODE field. Resulting table contains GRIDKECH field for spatial analysis.
species*_table2_sum	Data from species*_table2_grid summarized by GRIDKECH and NUM_FISH fields.
species*_table3	All species* data based on selection of unique NODC code from table3_reported .
species*_table3_grid	Data from species*_table3 joined to i5g_pr table by the ID_CODE field. Resulting table contains GRIDKECH field for spatial analysis.
species*_table3_sum	Data from species*_table3_grid summarized by GRIDKECH and FSHINSP fields.

Table 10: Shapefiles created for GIS analysis

Shapefile Name	*Description*
grid_albers	Resulting grid from the merge of NEFNgrid, NWETgird, SEETgrid, and SWETgrid shapefiles in the Albers Conical Equal-Area projection.
grid_allfish	Data from table2_grid and table3_grid joined to GRIDKECH field in the grid_albers shapefile. All cells completely contained by the boundary of Puerto Rico were removed from the shapefile so only cells located offshore of Puerto Rico were used in the spatial analysis.
grid_only_fish	Shapefile that contains records for only the cells with NUM_FISH and FSHINSP values greater than 1.
centroids_allfish	Shapefile that contains centroids for all cells in the grid_allfish

	shapefile.
centroids_only_fish	Shapefile that contains centroids for all cells in the grid_only_fish shapefile.

Table 11: GIS shapefiles created from the statistical analysis

Shapefile Name	*Description*
Gi_4900	Results of the Getis-Ord Gi* analysis with a distance band of 4,900 m (one cell).
Gi_9800	Results of the Getis-Ord Gi* analysis with a distance band of 9,800 m (two cells).
Gi_14700	Results of the Getis-Ord Gi* analysis with a distance band of 14,700 m (three cells).
Gi_19800	Results of the Getis-Ord Gi* analysis with a distance band of 19,800 m (four cells).
LMI_4900	Results of the Anselin Local Morans I analysis with a distance band of 4,900 m (one cell).
LMI_9800	Results of the Anselin Local Morans I analysis with a distance band of 9,800 m (two cells).
LMI_14700	Results of the Anselin Local Morans I analysis with a distance band of 14,700 m (three cells).
LMI_19800	Results of the Anselin Local Morans I analysis with a distance band of 19,800 m (four cells).
dolphin_GI_4900	Results of the Getis-Ord Gi* analysis on dolphin species with a distance band of 4,900 m (one cell).
dolphin_GI_9800	Results of the Getis-Ord Gi* analysis on dolphin species with a distance band of 9,800 m (two cells).
dolphin_GI_14700	Results of the Getis-Ord Gi* analysis on dolphin species with a distance band of 14,700 m (three cells).
dolphin_GI_19800	Results of the Getis-Ord Gi* analysis on dolphin species with a

	distance band of 19,800 m (four cells).
dolphin_LMI_4900	Results of the Anselin Local Morans I analysis on dolphin species with a distance band of 4,900 m (one cell).
dolphin_LMI_9800	Results of the Anselin Local Morans I analysis on dolphin species with a distance band of 9,800 m (two cells).
dolphin_LMI_14700	Results of the Anselin Local Morans I analysis on dolphin species with a distance band of 14,700 m (two cells).
dolphin_LMI_19800	Results of the Anselin Local Morans I analysis on dolphin species with a distance band of 19,800 m (four cells).

Table 12: Description of data fields

Field	Field Type	Description	Shapfiles and tables
ET_ID	Long Integer	Grid cell identification number	centroids_allfish, centroids_only_fish, grid_albers, grid_allfish, grid_onlyfish, species*_grid, all statistical shapefiles
GRIDKECH2	Short Integer	Spatial attribute used to join type 2 data to type 5 data. Represents the grid cell where the most fish were hooked for type 2 data.	centroids_allfish, centroids_only_fish, table2_sum, species*_grid, all statistical shapefiles
CNT_GRID_2	Short Integer	Numeric count of the GRIDKECH2 field. Indicates number of respondents for type 2 data with GRIDKECH values.	centroids_allfish, centroids_only_fish, table2_sum, species*_grid, all statistical shapefiles
NUM_FISH	Short Integer	Number of fish reported as hooked. Derived from type 2 data.	centroids_allfish, centroids_only_fish, table2_sum, species*_grid, all statistical shapefiles
GRIDKECH3	Short Integer	Spatial attribute used to join type 3 data to type 5 data. Represents the grid cell where the most fish where hooked for type 3 data.	centroids_allfish, centroids_only_fish, table3_sum, species*_grid, all statistical shapefiles

Field	Field Type	Description	Shapfiles and tables
CNT_GRID_3	Short Integer	Numeric count of the GRIDKECH3 field. Indicates number of respondents for type 3 data with GRIDKECH values.	centroids_allfish, centroids_only_fish, table3_sum, species*_grid, all statistical shapefiles
FSHINSP	Short Integer	Number of fish reported as hooked. Derived from type 3 data.	centroids_allfish, centroids_only_fish, table3_sum, species*_grid, all statistical shapefiles
ALL_FISH	Short Integer	Total catch value created by adding type 2 NUM_FISH data and type 3 FSHINSP data.	centroids_allfish, centroids_only_fish, grid_allfish, grid_onlyfish, species*_grid, all statistical shapefiles
AREA	Double	Area calculation for each grid cell measured in meters.	grid_allfish, grid_only_fish, species*_grid, all statistical shapefiles
ACRES	Double	Area calculation for each grid cell measured in meters.	grid_allfish, grid_only_fish, species*_grid, all statistical shapefiles

Field	Field Type	Description	Shapefiles
Gi4900, Gi9800, Gi14700, and Gi19800	Floating integer	Statistical output from the Getis-Ord Gi* analysis. The output of the Gi function is a z score for each feature. The z score represents the statistical significance of clustering for a specified distance.	All Gi shapefiles.
LMi4900, LMi9800, LMi14700, and LMi19800	Floating integer	Statistical output from the Anselin Local Moran's I analysis. The output of the Anselin Local Moran's I function is an index value. If the I index value is positive, then that feature has values similar to neighboring features' values. If the I index value is negative, then that feature is quite different from neighboring values.	All LMi shapefiles
LMz4900, LMz9800, LMz14700, and LMz19800	Floating integer	The z score resulting from the Anselin Local Moran's I analysis. The z score The z score represents the statistical significance of the Local Moran's I index value. It, in effect, indicates whether the apparent similarity (or dissimilarity) in values between the feature and its neighbors is greater than one would expect simply by chance. A high positive z score for a feature indicates that the surrounding values are similar values (either high or low). So, a group of adjacent features having high z scores indicates a cluster of similar values. A negative z score for a feature indicates the feature is surrounded by dissimilar values—that is, if a feature gets a negative z score, it has a different value than its neighbors (a high value relative to a neighborhood that has low values or a low value relative to a neighborhood that has high values). (ESRI, 2005).	All LMi shapefiles

Works Cited

Dungan, J., Perry, J., Dale, M., Legendre, P., Citron-Poutsy, S., Fortin, M., Jakomulska, A., Miriti, M. and Rosenberg, M. (2002). A Balanced View of Scale in Spatial Statistical Analysis. *Ecography* 25(5), 626-640.

Gotway, C. and Young, L. (2002). Combining Incompatible Spatial Data. *Journal of the American Statisitical Association*, (97(458) 632-648.

Holden, H., Derksen, C., and LeDrew, E. (2000). Coral Reef Ecosystem Change Detection Based on Spatial Autocorrelation of Multispectral Satellite Data. Retrieved November 1, 2005 from the GISDevelopment.Net Website: http://www.gisdevelopment.net/aars/acres/2000/ts3/cost006pf.htm.

Mitchell, A. (2005). *The ESRI Guide to GIS Analysis Volume 2: Spatial Measurements & Statistics*. United States: ESRI Press.

Svancara, L., Garton, E., Chang, K., Scott, J., Zager, P., and Gratson, M. (2002). The Inherent Aggravation of Aggregation: An Example With Elk Aerial Survey Data. *Journal of Wildlife Management*, 66(3), 776-787.

Wong, D. and Lee, J. (2005). *Statistical Analysis of Geographic Information With ArcView GIS and ArcGIS*. United States: John Wiley & Sons, Inc.