2020

Utilization of machine learning algorithms to support retail chain store location decisions

Petr Grin
University of Northern Iowa

UTILIZATION OF MACHINE LEARNING ALGORITHMS TO SUPPORT RETAIL CHAIN STORE LOCATION DECISIONS

An Abstract of a Thesis

Submitted

in Partial Fulfillment

of the Requirements for the Degree

Master of Arts

Petr Grin

University of Northern Iowa

July 2020

ABSTRACT

Businesses use GIS software to build spatial business models to have possibility to analyze customers, competitors or markets not only in terms of finance, but also in terms of their behavior in space. Specifically, geomarketing methods are used to maximize profit when searching for places to open new factories, shops, restaurants, or expand chain of cafes. The main geomarketing approach is to identify the optimal location based on socio-economic data, as well as on the criteria that are necessary for this type of business (suitability model).

Traditionally the Maxent model, a type of an ecological niche model (ENM), is used to forecast the distribution of different species of animals or plants in biological science. This machine-learning algorithm uses data from points where the phenomenon in question has already been found and selects locations with similar characteristics. Thus, the location is not determined by the expert choice, based on the conditions and factors that are perceived as necessary or beneficial, but is identified on the basis of the machine learning algorithms driven by statistical assessment of existing locational information for a given species. . However, his modeling approach has applications beyond species distribution and has already been used to examine other phenomena, such as optimal locations of wind turbines.

This study tests the opportunities of using Maxent for geomarketing applications, specifically determining the 'best' locations for coffee chain restaurants using Biggby locations in Michigan State as an example. The results indicate that Maxet and EMN principles could be potentially useful for geomarketing applications. The resultant model used a very limited number of variables but was able to demonstrate in principle the utility of the method. Also, 2 cell scales were used to compare how different the results of the models can be when using different scales. The main variables that were used in the study are based on those indicators that the Biggby coffee shop network uses directly to find new locations to expand its chain. The main limitations of this analysis are the lack of contact and support from Biggby, as well as the inaccessibility of certain types of data, as their traffic intensity. Another limitation in the work was the difficulty in processing a large amount of data, which was formed due to the fact that initially the analysis was to be applied to 4 states.

UTILIZATION OF MACHINE LEARNING ALGORITHMS TO SUPPORT RETAIL CHAIN STORE LOCATION DECISIONS

A Thesis

Submitted

in Partial Fulfillment

of the Requirements for the Degree

Master of Arts

Petr Grin

University of Northern Iowa

July 2020

This Study by: Petr Grin

Entitled: UTILIZATION OF MACHINE LEARNING ALGORITHMS TO SUPPORT RETAIL CHAIN STORE LOCATION DECISION

has been approved as meeting the thesis requirement for the Degree of Master of Arts

_____	_____
Date	Dr. Andrey N. Petrov, Chair, Thesis Committee
_____	_____
Date	Dr. Alex Oberle, Thesis Committee Member
_____	_____
Date	Dr. Arti Mann, Thesis Committee Member
_____	_____
Date	Dr. Jennifer Waldron, Dean, Graduate College

TABLE OF CONTENTS

PAGE

LIST OF FIGURES

PAGE

LIST OF TABLES

PAGE

LIST OF PLOTS

PAGE

LIST OF SCRIPTS

PAGE

CHAPTER 1

INTRODUCTION

Location intelligence software such as Geographic Information Systems (GIS) is used in different areas of human activity such as mining, human influence on environment, ecology, climate change (Sherrouse, Clement, and Semmens 2011). Location analysis is also useful for urban development and planning. GIS helps to keep, manipulate, and analyze different data types in the city, such as physical, social and economic data. Planners can then apply analytical and mapping functionality of GIS to examine the existing social economic, infrastructure or ecological situation in the city (Yeh 1999). It is already an integral part of designing the traffic flows and supply chains where attribute analysis in GIS is used to provide an optimal solution to manage costs of logistics (Irizarry, Karan, and Jalaei 2013). City authorities understand that urban environment needs to be planned, and different areas should perform certain functions. For example, the New York authorities created an attractive business environment for more efficient use of urban potential in the city and how we can apply this method in Europe region (Ward 2005).

Commercial organizations also use location analysis to increase revenue from each opened point of sale or production. A tightening race to attract customer attention in business and services planning, along with advances in GIS and spatial analysis techniques, have led to the promotion

of the use of GIS in the area of business and service planning (Hernández and Bennison 2000). GIS allows businesses to build long-term models which take into account many factors that influence their strategy and decision making (Pick 2008). GIS can be used not just for location and planning analysis, but also for solving other retail tasks such as, merchandising, category management, marketing communications and relationship marketing.

A number of technologies are now widely available and utilized to analyze the spatial structure of retail activities with location data at a microscale. These include application of methods such as Probability Density Function (PDF), Decision Support Systems (DSS), Spatial Interaction Models, Network Huff Model, Analysis of Variance (ANOVA) (Byrom 2005), MATISSE ("Matching Algorithm, A Technique for Industrial Site Selection and Evaluation"), and RASTT (Retail Aggregate Space Time Trip Model) (Baker 2002), Suitability modeling (Johnston and Graham 2013) and others (Sugumaran and Degroote 2011).

However, these models do not capitalize on the fact that a company may already have a number of points of sale which can be used to understand location requirements for new facilities through data mining and analysis. This opens a possibility to utilize machine learning algorithms and tools, such as logistic regression or decision tree algorithms. One of such instruments which can be used is Maxent. The

Maxent model was originally intended to predict the habitats of various species of animals, based on existing locations REF. This method is called ecological niche modeling, which means that we can find the same kind of species in the area with the similar values of environment or same habitat. The algorithm uses data from existing points and selects locations with similar characteristics. Thus, the location is not determined by the choice made by experts based on the necessary conditions and factors, but on the basis of the 'machine learned' conditions known for those points in which these species of animals or plants. The Maxent model has already been used not only to predict the location of biological species (Phillips et al. 2017), but also to find most suitable places for human structures, such as wind turbines (Petrov and Wessling 2015). The model is based on the principle of ecological niche modeling. Thus, the model can perceive variables of completely different kinds. The algorithm uses the distribution of maximum entropy. Maxent output is a surface that indicates the predicted probability that the conditions at this point fit (Phillips, Anderson, and Schapire 2006).

The random forest classifier consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector (Breiman 1999).

Thus, the logistic regression, which is part of the Maxent model, is a tool for dividing input data into two classes, based on the known indicators of variables. The main parameter of logistic regression is the parameter C, which determines how strictly the model makes the classification. In this way, the degree of training and the rigor of the model can be adjusted. Algorithm random forest, in turn, is also a classifier, but based on a decision tree, where the final classification depends on a sequential analysis of the variables and their influence on the final result. In addition, the forest does not control with what parameter the classification begins. The main parameters of this model is determining how many decision branches will be in the model and the number of divisions in each branch.

1.1 Research Goal

The goal of this research is to determine whether it is possible to use the ecological-niche and decision tree modeling in business location analysis, and specifically for locational decision regarding new store placement within a retail food chain. Evaluate how these two algorithm works on different scale. And how we can use them to solve strategic development questions for retail chains.

1.2 Research Questions

Based on the goal of research this thesis addresses the following research questions:

1. How can machine learning be used to develop predictive models of restaurant chain locations?
2. What is the accuracy of machine learning algorithms for locating shops in a restaurant chain?
3. What are the advantages and disadvantages of machine learning in comparison to decision-tree algorithms for modeling restaurant chain locations?

1.3 Research Objectives

To solve posed research questions I have to do next research objectives:

1. Calibrate a machine learning algorithm to develop a predictive model of potential store locations using variables and specifications found in retail chains business modeling.
2. Perform machine learning-based modeling of potential store locations using Maxent at different scales.
3. Compare the outputs and performance of the ecological niche modeling, and random forest machine learning algorithms.

1.4 Significance of Anticipated Results

Results of this research can be used in business decision making processes, business intelligence, and geomarketing specifically. Business

owners can use these algorithms to understand how it is possible to open new points of sale in the new areas. Results can be used to determine usefulness of machine learning in studying the development strategy of competitors.

CHAPTER 2

LITERATURE REVIEW

2.1 Using GIS in Business and Decision Making

Growing business demand for better customer satisfaction leads to different technical methods of analysis for decision making. Distinct business areas come to the fact that space plays an essential role in success, and GIS is an essential tool in making business development decisions (Longley and Clarke 1995). GIS is recognized by many retail companies and organizations as Walmart, McDonald's and Starbucks. Business geographic courses are now incorporated in business curricula and business geography analytics has become a recognized niche of study (Jones and Hernandez 2004). The vast literature reveals the power of GIS as a tool for decision making and product promotion (Sugumaran and Degroote 2011).

Clarke (1998) wrote that in an era when many companies are striving to cut costs, any competitive advantage can be played for profit. Thus, the search for the optimal geographical location becomes an integral element of the business processes. For example Canadian small retailers groups felt the pressures of big international American companies in the early 2000s. They tried to find new ways to keep their share of the market and started using spatial models to determine more profitable locations (Hernandez and Biasiotto 2001).

Averse mentioned that the location decision making usually made on the basis of the classic data set, which does the spatial aspect. However, today many retailers are starting to use spatial big data for decision-making, which they have been collecting for quite some time, but retailers simply did not know what to do with them. Retailers are starting to buy more and more software, which is geared towards analyzing spatial data. The amount of data itself is becoming bigger, so we can say that retailers are beginning to analyze big spatial data (Aversa 2019).

In his work, France described that in the near future, cooperation between the academic environment and marketing practices will only increase. First of all, due to the increase in the number of graduate students in the field of big data analytics. The number of data analysis courses at marketing departments is also growing. It was also noted that more and more scientists want to interact with the industry (France and Ghose 2019). Geography is becoming applied science and takes on the traits of the craft between consulting, IT and big data analysis (Clarke 1998). Many studies concluded that the use of geo-information technologies makes it possible to further apply GIS in the area of developing services and planning business processes and making decisions. Compiling data that is accumulated in companies and employing spatial analysis techniques provide a great competitive advantage for companies. However, the models that are used in GIS need

to be critically analyzed in respect to the characteristics and variables that they use (Longley and Clarke 1995).

Using GIS with other analytical software can solve problems in several ways, such as location studies in retail, restaurant business or manufacturing. Thus, the models used in GIS can be adapted to different tasks and needs. Spatial modeling and big data are used to analyze and forecast supply and demand REF. However the main topic this research is the use of machine learning algorithms and the comparison of their results with the prediction of suitability modeling. The result of the study will be an analysis of whether it is possible to use machine learning for decision-making and at what scale the results are more suitable (Birkin, Clarke, and Clarke 2002).

2.2 Retail Location Analysis Using GIS Methods

Retail location analysis in GIS employs various techniques to study optimal locations for restaurants chains. Among these methods are:: Probability Density Function (Sadahiro 2001), Spatial Interaction Models (Lao 1993), Huff Model (Liu 2012), Network Huff Model (Okabe and Okunuki 2001), Voronoi diagrams (Okabe et al. 2009), Decision Support Systems (Hess, Rubin, and West 2004), Matching Algorithm, A Technique for Industrial Site Selection and Evaluation (Witlox 2003), and Retail Aggregate Space Time Trip Model (Baker 2002), among others.

Probability Density Function

The probability density function characterizes, the density with which the values of a random variable at a given point are distributed. Sadahiro (2001) used this model to analyze spatial structure of retailing. He wrote that the probability density function measures the degree of agglomeration, enables classifying spatial patterns of store location, the relationship between the size and function of retail agglomerations, and determining the spatial structure of retail agglomeration (Sadahiro 2001).

Spatial Interaction Models

SIM models are used to find better locations for retail shops provided that the system allocations (cash flows or customer trips) will be following some sorts of spatial interaction rules. Huff model can be a part of SIM models. The SIM model can include several types of variables: demand, supply, travel cost, spatial structure and facilities. Lao Young (1993) wrote that Spatial Interaction Models use three approaches: the cost minimizing approach, the benefits maximizing approach, and the entropy maximizing approach. Each approach represents a unique perspective on how new retail outlets should be chosen in the context of many complicated issues - demand versus supply, cost versus benefit, and competition versus cooperation (Lao 1993).

Network Huff Model

The network Huff model is formulated on a network with the shortest-path distance as an extension of the ordinary Huff model. Okable and Okunuki (2001) derived formula for estimating the demand i.e. that computes the demand on a network. The proposed method is easily modifiable to take more detailed spatial factors into account. First, a physical route distance can be replaced with a travel time distance. Second, we can easily treat psychological distances, such as the `intersection distance' (Sadalla and Staplin 1980), `corner distance' (Sadalla and Magel 1980), `slope distance' (Okabe, Aoki, and Hamamoto 1986), or more generally the cognitive distance. Third, we assume that the attractiveness, ai, of a store is represented only by the floor area, but we may consider ai as a function of many variables which are discussed in the sophisticated Huff models referred to in the introduction.

Decision Support Systems

Retailers are increasingly developing GIS as DSS for sales promotion and long-term strategic decision-making. GIS merges endogenous databases by retailers and the exogenous data sources to introduce retail decision-making and support systems implementation. As an example, the examination of the experiences of some of the UK based retailers reflecting GIS implementation in retail location analysis shows a highly organized

series of process management that has resulted from this application (Hess, Rubin, and West 2004). Sugumaran and Degroote wrote about history of Spatila Decision Support Systems development. “SDSS had an exponential growth in supporting commercial, governmental, and academic decision-making processes for situations more complex than just deciding where to have dinner. Issues such as environmental management, land use planning, transportation design, commercial or public welfare service provision, and emergency/hazard management cut across administrative, institutional, and stakeholder settings, which are couched within a complicated spatial matrix” (Sugumaran and Degroote 2011).

Matching Algorithm, a Technique for Industrial Site Selection and Evaluation

MATISSE is a knowledge-based decision support system (KBDSS) based on decision tables that can be used by industrial decision-makers and planners to assess the suitability of potential sites (Witlox 2003). Witlox explains how a relational approach to the modeling of the site suitability concept can be implemented and tried to find all possible locations that meet the spatial production requirements based on the organizational characteristics of the firm. The growing interest of urban geographers and economic geographers in applying KBS, DSS and integrated system has been largely attributed to the development of computer systems.

Computers are able to store, organize and process enormous amount of data as well as make possible the availability and accessibility of the domain-specific knowledge underlying the spatial problem.

Maxent

Maxent is an open source software program from open Modeller. In its traditional use, Maxent estimates a species distribution based on known occurrence locations of a species and landscape variables over the entire study area (Phillips, Anderson, and Schapire 2006). The advantage of the Maxent algorithm is that it is able to predict the probability that conditions are suitable for locating an object using presence-only data, i.e. known locations of existing objects. Maxent has traditionally been used to model species distributions and has also been utilized to model abiotic phenomena. Benito and Peñas de Giles (Benito and Peñas 2008) used Maxent to create suitability maps for greenhouses in Spain. In another study, Parisien and Moritz used Maxent to rate locations most at risk for wildfires in California (Parisien and Moritz 2009).

CHAPTER 3

METHODOLOGY

3.1 Machine Learning Algorithms in Location Analysis

One of the common ways to find locations for a business is to directly calculate traffic, human traffic, and other necessary parameters at the site of the planned opening REF. Today, counting traffic on the street is an effective method to understand the number of potential customers of a restaurant or cafe. The number of passers-by can vary greatly by time of day, days of the week, so there are many methods for calculating passers-by. One of the most interesting methods is counting passers-by using video recording (Fujisawa 2013).

The method of suitability location analysis, which is carried out by applying GIS technologies, is also widespread (Sugumaran and Degroote 2011). The model is based on the spatial parameters necessary for a business. The result is obtained by combining all the parameters with the addition of various coefficients. The coefficients are selected by the researcher himself, which introduces the influence of the human factor into the model, which sometimes can lead to a decrease in the accuracy of the model.

In my research, I want to explore the possibility of using machine learning in business applications, as part of the search for new locations. The main differences of my research will be:

1. The use of existing locations as values for teaching my model, and external factors will play the role of the environment in which existing points are located. New points should appear in locations with similar environment settings. Unlike previous studies, I do not set the parameter or weight for each variable in the spread, but the algorithm classifies each variable by the weight of the contribution to the location.
2. Another difference from similar studies will be the use of variables that are built into the business model by the company itself. Thus, my research is getting closer to using it in practice.

One of the modern approaches to data analysis is the use of logistic regression. Logistic regression is used in many sectors of the economy, from the banking sector, to car production REF. In banking, the use of logistic regression in the process of scoring is widespread, which increases the efficiency of banking services (Bensic, Sarlija, and Zekic-Susac 2005). Based on the existing data (variable) for each row in the database, as well as on the already known results for these rows, we can predict what the result will be for new rows in the database.

A nonlinear logistic has having the following form:

$$Y = \frac{1}{1 + e^{y}}, y = C_0 + \sum_{i=1}^{n} c_i X_i$$

Where X_i 1 < i < n, represent the set of individual variables, c_i is the coefficient of the i −th variable, and Y is the dependent variable. Since Y falls between 0 and 1, it is usually interpreted as the probability of a class outcome (Tam and Kiang 1992). The advantage of logistic regression is that, through the addition of an appropriate link function to the usual linear regression model, the variables may be either continuous or discrete, or any combination of both types and they do not necessarily have normal distributions (Lee 2005).

Logistic regression has also established itself in the geographical sciences. For example, machine learning was used to analyze the best locations for wind turbines (Petrov and Wessling 2015). To determine the location for the new turbines, the existing locations of the wind turbines were used, as well as the data located at these points. The following data were used as variables: wind speed, slope steepness, distance from roads and others. The study showed that the use of logistic regression gives a working result, and its accuracy is 85%.

Of course, the application of logistic regression in geography has its drawbacks. So, different data have different accuracy. Some data may be presented only at the state or county level, while others may have the accuracy of a track or even district. Thus, correlating data is quite difficult and the result becomes less accurate.

Machine learning

According to Tom M. Mitchell "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." In-plane words, machine learning is an approach where we try to give intelligence to the machine so that it can do a specific task more like a human. (Mitchell 1997).

As a classification of machine learning, you can divide it into a supervised and unsupervised. Supervised classification give us possibility to input parameters X and output parameters Y. So, our role is to teach our algorithm give us right outputs if we will have new inputs. To this class we can include classifications and regressions. Unsupervised classification needs only inputs and doesn't ask us to provide outputs to teach itself. To this class we can include clustering and associations.

In the field of machine learning, semi-supervised learning occupies the middle ground, between supervised learning (in which all training examples are labeled) and unsupervised learning (in which no label data are given). Interest in Semisupervised learning has increased in recent years, particularly because of application domains in which unlabeled data are plentiful, such as images, text, and bioinformatics. Basically, it is the mixture of labeled and unlabeled data. It is useful in many ways. Labeling a huge amount of data is time-consuming and expensive. Add to

that, it also can impose human biases on the model. Thus, it means that supervised learning improves the accuracy of the model. The little bit of labeled data is the starting point for the algorithm and kind of a clue to start its labeling and to label a large amount of data. This approach is widely used in genetic sequencing, web page classification and so on.

Reinforcement learning is the last branch among the 4 branches of machine learning. It allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behavior; this is known as the reinforcement signal. This behavior can be learned once and for all, or keep on adapting as time goes by. If the problem is modeled with care, some Reinforcement Learning algorithms can converge to the global optimum; this is the ideal behavior that maximizes the reward. This automated learning scheme implies that there is little need for a human expert who knows about the domain of application. Much less time will be spent designing a solution, since there is no need for hand-crafting complex sets of rules as with Expert Systems, and all that is required is someone familiar with Reinforcement Learning (Shihab et al. 2018).

One of the main advantages of machine learning is that the more data models have, the more accurate their result becomes. Thus, the model learns and each time cross-checks itself with existing results.

However, it is necessary to carefully select the variables for the models, since too many variables can lead to an incorrect result.

3.2 Maxent

Maxent is a universal method that gives you the possibility to make predictions or draw conclusions from information which is not complete. Originally this method comes from statistical mechanics and it is actively used in such areas as astronomy, image processing, optimization of portfolio, physics, signal analysis and species distribution (Jaynes 1957). I will use it as an approach for modeling cafes locations in Iowa, based on already existing cafes locations in Michigan.

The Maxent idea is to evaluate probability distribution by analyzing and finding maximum entropy probability distribution. Finding the probability distribution of maximum entropy, taking into account a number of constraints that represent our incomplete information about the target distribution. The available information about distribution is usually presented as a set of variables which are called “objects”, and the limitations are that the expected value of each object should correspond to its empirical average (average value for a set of samples) (Phillips and Dudık 2007).

To distribute the probability, the average value of all variables at the scene is used, which serves as the target value. Of all the distributions, only the distribution of maximum entropy is selected. It refers to the most

uniform distribution, which increases uncertainty. This approach produces a more natural distribution of phenomena. The distribution that Maxent generates is characterized by maximum entropy. (Phillips and Dudık 2007). Maxent uses an exponential model and empirically derived functions of environmental variables that are subject to a sequential update algorithm that adjusts the coefficients to achieve the desired distribution. The algorithm is deterministic and guaranteed to converge to the maximum probability distribution of entropy (Petrov and Wessling 2015).

When Maxent is used to model the distribution of cafes using presence-only data, the pixels of the study area constitute the space in which the probability distribution of Maxent is determined, the pixels with known records of store locations constitute sampling points with characteristics, such as: income, distance from competitors and others (Phillips, Anderson, and Schapire 2006). In my case, Maxent is used as a tool for analysis and modeling of the ecological requirements of the Biggby caffes, together with a set of environmental variables that describe some factors that can affect the suitability of the environment for the species (Root 1988). Each entry is a latitude-longitude pair, indicating the place where the existing cafe is located. Similar records can be found in museums, where exhibits are presented along with a description of the geographic location and surrounding conditions (Ponder et al. 2001). All

environmental variables in the GIS format belong to the same geographical area - the study area, which was divided into a grid of pixels. The task of the modeling method is to predict “the ecological suitability” of cafes depending on the given environmental variables.

The niche-based model is an approximation of the ecological niche of the species in the studied environmental measurements. The fundamental niche of a species consists of a set of all conditions ensuring its long-term survival, while its realized niche is the subgroup of the fundamental niche that it actually occupies (Hutchinson 1957). A realized niche of a species may be smaller than its main niche due to human influence, biotic interactions (for example, interspecific competition, predation) or geographical barriers that prevented settlement and colonization; such factors can impede the habitat of the species (or even its occurrence) in conditions that encompass its full ecological potential (Pulliam 2000). Here we assume that the sites of origin originate from the original habitat, and not from the habitat that may contain the species without the conditions necessary to maintain the population without immigration; this assumption is less realistic for very fuzzy taxa (Pulliam 2000). Thus, by definition, environmental conditions in places of origin are samples from a niche realized. Thus, a niche-based model is an approximation of a realized niche of a species in the study area and environmental aspects.

3.2 Decision Tree and Random Forest

Decision tree learning is one of the predictive modeling approaches used in statistics, data mining and machine learning. Shalev-Shwartz in his book "Understanding Machine Learning: From Theory to Algorithms" described a decision tree as follows: A decision tree is a predictor, $h : X \rightarrow Y$, that predicts the label associated with an instance x by traveling from a root node of a tree to a leaf. At each node on the root-to-leaf path, the successor child is chosen on the basis of a splitting of the input space. Usually, the splitting is based on one of the features of x or on a predefined set of splitting rules. A leaf contains a specific label. A popular splitting rule at internal nodes of the tree is based on thresholding the value of a single feature. That is, we move to the right or left child of the node on the basis of $1_{[x_i < \theta]}$, where $i \in [d]$ is the index of the relevant feature and $\theta \in$ R is the threshold. In such cases, we can think of a decision tree as a splitting of the instance space, $= R^d$, into cells, where each leaf of the tree corresponds to one cell. A general framework for growing a decision tree is as follows. We start with a tree with a single leaf (the root) and assign this leaf a label according to a majority vote among all labels over the training set. We now perform a series of iterations. On each iteration, we examine the effect of splitting a single leaf. We define some "gain" measure that quantifies the improvement due to this split. Then, among all possible

splits, we either choose the one that maximizes the gain and perform it, or choose not to split the leaf at all (Shalev-Shwartz and Ben-David 2014, 2).

In the decision tree method, sequential trees give extra weight to points incorrectly predicted by earlier predictors. At the end, a weighted vote is taken for the forecast. In batching, consecutive trees are independent of earlier trees — each one is independently constructed using the initial load of the dataset. In the end, a simple majority of the votes is taken for the forecast.

Breiman (1999) proposed random forests that add an extra layer of randomness to packaging. In addition to constructing each tree using different boot data samples, random forests change how classification or regression trees are built. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best of the subset of predictors randomly selected in that node. This somewhat controversial strategy is very good compared to many other classifiers, including discriminant analysis, reference vector machines and neural networks, and is resistant to refitting (Breiman 2001). In addition, it is very convenient in the sense that it has only two parameters (the number of variables in a random subset at each node and the number of trees in the forest), and are usually not very sensitive to their values.

Thus, in order to avoid incorrect classification by one decision tree, in my research I will take the random forest method. As inputs (x) for random

forest will be variables there will be socio-economic data, data on the location of competitors and also data on the location of infrastructure, which increases the attractiveness of the place for opening a cafe. Such variables include parking lots and intersections. As outputs (y) I will use already existing Biggby cafe points locations.

3.3 Study Area

The study area of my research is the State of Iowa. According to the United States Census Bureau estimated population of Iowa should be 3 167 997, that is more than 4% bigger than in 2010 (the year of last Census). Area of Iowa is 145,746 km^2 the 26th place among all US states. Iowa borders along the Mississippi River in the east and the Missouri River and Big Sue River in the west. Northern border is a line along 43 degree and 30 minutes of north latitude. The south border is the line along 40 degrees 35 minutes north of north latitude.

According to the United States Census Bureau urban population of Iowa is 1 950 256 people and 1 096 099 are rural. That means that 64% of the total population in Iowa lives in cities. The biggest city by population in Iowa by the United States Census Bureau is Des Moines 217,891 people, the second one is Cedar Rapids with 135,502 people and Davenport with 101,965.

City	Population
Des Moines	217,891
Cedar Rapids	135,502
Davenport	101,965
Sioux City	81,382
Iowa City	78,440
Ankeny	72,820
West Des Moines	68,619
Ames	67,962
Waterloo	67,228
Council Bluffs	62,289

Table 1. Top 10 cities by population in Iowa 2020

3.4 Iowa Economic Characteristics

Iowa always presents itself as a farming state. However, agriculture occupies a small percentage of the state economy. Today, the Iowa economy is highly diversified and has such components as manufacturing, biotechnology, finance and insurance services (Iowa Workforce Development 2004).

Row Labels	Sum of GDP (Millions of current dollars)	Sum of GDP (Millions of current dollars)
Manufacturing	18.59%	141,034
Finance and insurance	14.02%	106414.2
Government	11.33%	85950.7
Real estate and rental and leasing	10.23%	77630.1
Health care and social assistance	6.59%	50029.6
Wholesale trade	6.15%	46648.8
Retail trade	5.24%	39758.2
Construction	4.21%	31917
Agriculture, forestry, fishing, and hunting	3.51%	26654.6
Professional, scientific, and technical services	3.43%	26018.8
Grand Total	100.00%	758815.7

Table 2. Top 10 Iowa Gross Domestic Product by the year 2018 (data.iowa.gov)

According to the State of Iowa's data portal agriculture sector only takes 3.51% of Iowa State GDP with 26654.6 millions of dollars. The largest industry in terms of GDP in 2018 was manufacturing with 141034 millions of dollars of GDP (State of Iowa data portal 2019). The main industries are the food industry, heavy equipment manufacturing, as well

as the production of chemicals for agriculture. Sixteen percent of Iowa's workforce is dedicated to manufacturing (Iowa Workforce Development 2004).

Food manufacturing is one of the largest sectors of the economy in Iowa. In addition to food processing, this segment also includes the production of chemical products and electrical equipment, as well as the production of machinery used in agriculture. In addition, Iowa has a large production of non-food products. One of the largest companies is John Deere with the revenue of $29,738 million in 2017 (“Annual Report John Deere” 2017).

Finance and insurance sector takes 14.02 % of Iowa GDP with 106414.2 million dollars. In 2017, this sector accounted for 6.2 percent of the total employed population in the private and public sectors of Iowa. Salaries in this sector, the average annual salary for 2017 amounted to 75,662 dollars. This is 64.2 percent more than the average state salary, which amounted to $ 46,073 dollars. This sector includes companies whose main activity is participation in financial transactions (transactions related to the creation, liquidation or change of ownership of financial assets) as well as in facilitating financial transactions. There were 6,610 finance and Insurance locations in Iowa in 2107. Biggest employers in this sector were: EMC Insurance Co, Principal Financial Group Inc,

Transamerica Life Insurance Co, Wellmark Inc and Wells Fargo Bank (Iowa Workforce Development, Finance & Insurance Iowa Industry Profile 2018).

Real estate and rental and leasing sector takes 10.23% of Iowa GDP with 77630.1 million dollars. In 2017, 14,704 people worked in the real estate, rental and leasing sector in Iowa. This number is 1% of all employees in the state. This industry suffered quite a lot during the crisis, and employment decreased by 13.4% from 2005 to 2010. However, after 2010, employment in this sector grew by 14.4%. Salaries in this sector have risen significantly since 2008 and increased by 36.5 percent to $ 45,609. Now wages in this sector roughly correspond to the average Iowa wage. The biggest employers in this sector are: Conlin Properties, Duke Aerial Equipment, Hubbell Realty, Skeffington's Formal Wear and Vanguard Appraisals (Iowa Workforce Development, Real estate and leasing Iowa Industry Profile 2018).

Health care and social assistance sector take 6.59% of Iowa GDP with 50029.6 million dollars. In 2017, this sector was the largest in Iowa. 14.5 percent of all employed worked in this sector. From 2008 to 2017, this sector increased the number of employees by 12, 4%. Salaries increased by 20.9%. The crisis practically did not affect this industry, and the number of jobs has constantly increased over the past ten years. However, wages in this sector are below the state average by 7% and amounted to $ 44,649 in 2017. This industry includes companies that

provide medical and social assistance to individuals. These sectors of the economy include not only medical, but also social assistance (Iowa Workforce Development, Health care and social assistance Iowa Industry Profile 2018).

According to the Committee for Economic Development, the accommodation and food sector take 0, 72% of total Iowa GDP and got 5311 million dollars in 2018.

3.5 Selecting a Retail Chain for Modeling

The object of my research will be a network of coffee houses. In order to choose a set, the following criteria were adopted: the number of points in the network, the adaptability of the menu to the Midwest and Iowa, particular non-ethnic cuisine, and the availability of reviews on the Yelp social network.

European or American cuisine. When choosing a restaurant chain, the decision was made to exclude ethnic cuisine and focus on European or American cuisine. However, since many Chinese and Mexican dishes have long been included in the local culture, chains such as Panda Express and Taco Bell will not be ethnic restaurants.

Adaptability the menu for the Midwest and Iowa. Historically, the prevalent diet of Iowa takes origins from the German, Norwegian, British and Dutch diets of the 19th century immigrants and is characterized by a high content of meat and cereals (“USDA/NASS 2018 State Agriculture

Overview for Iowa" 2019). However, in addition a mix of Italian, Mexican and Chinese cuisines is very popular. Thus, we can say that these cuisines are common in the diet of Iowa residents. In addition, fast food (burgers, hot dogs and sandwiches) plays a large part.

As elsewhere, coffee is one of the most common drinks, so there are a large number of coffee houses in Iowa. Thus, we can say that for the Midwest and in particular, for Iowa, European cuisine with a greater share of meat dishes, as well as Chinese and Mexican cuisine, will be acceptable. Coffee shops are also in great demand.

The availability of reviews on the Yelp social network. This parameter is necessary in order to select a network which locations have continuous interaction with customers. The easier it is to find reviews about cafes in the Yelp network, the more representative is each of the cafes in this network. Therefore, reviews in this situation play the role of an indicator of the cafe's interaction with customers. To include a point in my analysis, it must have a rating above the median, if there is a rating higher than 3.0.

The number of stores in the network. The minimum number of points in the network is determined by the requirements of the Maxent algorithm. At its core, Maxent has logistic regression, which allows you to distribute input data into two groups. The study of van Proosdij says that the minimum number of points varies from 11 to 45, depending on the

prevalence of the species(Proosdij et al. 2016). Maxent's User Guide t states that there is no specific amount of data that needs to be used for the model to work properly. The manual notes that it is advisableto run the model several times with different amounts of test data in order to find a more optimal result (Young, Carter, and Evangelista 2011). Long (1997), based on his experience,suggests that maximum likelihood estimation including logistic regression with less 100 cases is “risky,” that 500 cases is generally “adequate,” and there should be at least 10 cases per predictor. Based on simulations, (Peduzzi et al. 1996), refine the 10:1 recommendation, stating that ten times the number of predictors, k, should take into account the proportion, p, of successes, n = 10k/p. Thus, for this research, it is better to take cafe networks with a large number of points for the possibility of checking the model. Thus, the number of points of existing points must exceed 100. In addition, it was necessary to be able to find the addresses of all points that the network has.

The number of points in the Midwest. This metric is important because our model must be applied to Iowa, which is part of the Midwest. Therefore, it is advisable that as the chain has stores in the Midwest, to capture similarities in socio-demographic indicators, natural factors, and development. This will give more relevant information about the location of the cafe, as in different regions its own characteristics. In the Midwest,

I include such states as: Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin.

The number of stores in Iowa. This study involves an analysis of the expansion of the coffee shop network in Iowa, so there should be no existing shops in the state. *The ratio of the number of stores in the Midwest to the total number of locations.* This indicator makes it possible to understand more precisely the concentration of this restaurant chain in the Midwest. For our model, it is important that more points of this network are in the Midwest.

Network coverage in the states. The more states inside the Midwest that are covered by a chain of coffee houses, the better. Thus, the sample points will be located in places with various values of variables, which will give a more detailed picture of where the coffee beans of the Biggby network can be opened.

Based on the above criteria, I have selected the following network cafes and restaurants:

Chain	Points	Points in Midwest	Points in Iowa	Points in Midwest/Points	Total coverage (states)	Non-ethnic restaurant	Applicable cafes model for Midwest	Possibility to find cafe rating on Yelp
Au bon pain	175	33	1	19%	26	+	+	100%
Biggby	250	239	0	96%	8	+	+	100%
Fuddruckers	144	15	0	10%	32	+	+	90%
Gloria Jeans	62	33	0	53%	23	+	+	81%
Jamba Juice	796	41	1	5%	35	+	+	97%
Peet's coffee & tea store	253	13	0	5%	11	+	+	91%
Pretzelmaker	183	34	6	19%	47	+	+	66%
Smoothie King	973	131	3	13%	32	+	+	95%
Jack in the box	2240	83	0	4%	21	+	+	98%
In-N-Out Burger	342	0	0	0%	7	+	+	88%
Cracker-Barral	654	101	3	15%	45	+	+	99%

Table 3. Table of selection criteria for a restaurant chain

Chain selection. To choose the most suitable restaurant chain, I decided to compare the indicators that I identified as important. However, it is impossible to simply compare their values with each other, so for comparison I had to standardize them or calculate Z score value (Greenacre and Primicerio 2008). To standardize the value, I use the following formula:

$$Z = \frac{X - \mu}{\sigma}$$

Where X is original value, μ is mean value and σ is standard deviation

Chain	Number of locations	Standardized
Au bon pain	175	-0.596
Biggby	250	-0.477
Fuddruckers	144	-0.645
Gloria Jeans	62	-0.774
Jamba Juice	796	0.386
Peet's coffee & tea store	253	-0.472
Pretzelmaker	183	-0.583
Smoothie King	973	0.665
Jack in the box	2240	2.667
In-N-Out Burger	342	-0.332
Cracker-Barral	654	0.161

Table 4. Number of points and standardized value

For each parameter, its standard deviation, average, was calculated, and after that, for each value, Z score was calculated.

Chain	Points in Midwest	Z-score
Au bon pain	33	-0.466
Biggby	239	2.469
Fuddruckers	15	-0.723
Gloria Jeans	33	-0.466
Jamba Juice	41	-0.352
Peet's coffee & tea store	13	-0.751
Pretzelmaker	34	-0.452
Smoothie King	131	0.930
Jack in the box	83	0.246
In-N-Out Burger	0	-0.936
Cracker-Barral	101	0.503

Table 5. Number of points in Midwest and z-score value

The assessment is based on the fact that it is important for the model that the number of points in the Midwest be as larger relative to the total number of points in the USA. This will give our model a more accurate result, as it will analyze the type of point arrangement closer to Iowa.

The assessment is based on the fact that for the model it is important that the number of points in Iowa is minimal for a cleaner experiment.

Chain	Points in Iowa	Z-score
Au bon pain	1	0.140
Biggby	0	0.651
Fuddruckers	0	0.651
Gloria Jeans	0	0.651
Jamba Juice	1	0.140
Peet's coffee & tea store	0	0.651
Pretzelmaker	6	-2.419
Smoothie King	3	-0.884
Jack in the box	0	0.651
In-N-Out Burger	0	0.651
Cracker-Barral	3	-0.884

Table 6. Number of points in Iowa and standardized value

For the number of points in Iowa Z score was multiplied by -1 because the smaller the number of dots, the better.

Chain	Points in Midwest/Point amount	Z-score
Au bon pain	19%	-0.103
Biggby	96%	2.602
Fuddruckers	10%	-0.401
Gloria Jeans	53%	1.108
Jamba Juice	5%	-0.586
Peet's coffee & tea store	5%	-0.587
Pretzelmaker	19%	-0.113
Smoothie King	13%	-0.293
Jack in the box	4%	-0.637
In-N-Out Burger	0%	-0.768
Cracker-Barral	15%	-0.223

Table 7. Number of points in Iowa and standardized value

The assessment is based on the fact that it is important for the model that the number of points in the Midwest be as large as possible for a more accurate model.

Chain	Total coverage (states)	Z-score
Au bon pain	26	-0.007
Biggby	8	-1.313
Fuddruckers	32	0.429
Gloria Jeans	23	-0.224
Jamba Juice	35	0.647
Peet's coffee & tea store	11	-1.095
Pretzelmaker	47	1.517
Smoothie King	32	0.429
Jack in the box	21	-0.369
In-N-Out Burger	7	-1.385
Cracker-Barral	45	1.372

Table 8. Number of sates covered by chain

The assessment is based on the fact that it is important for the model that the number of states that this restaurant chain covers is maximized.

Chain	Possibility to find cafe rating on Yelp	Z-score
Au bon pain	100%	0.834
Biggby	100%	0.834
Fuddruckers	90%	-0.095
Gloria Jeans	81%	-1.016
Jamba Juice	97%	0.522
Peet's coffee & tea store	91%	-0.035
Pretzelmaker	66%	-2.457
Smoothie King	95%	0.313
Jack in the box	98%	0.664
In-N-Out Burger	88%	-0.340
Cracker-Barral	99%	0.776

Table 9. Possibility to find cafes in Yelp

The assessment is based on the fact that it is important for the model that there should be as many points with ratings in Yelp as possible. To check the availability of the rating, a random selection was made for every third point in each restaurant chain. The ratio of points with rating and points without rating gives the percentage of reviews in Yelp. The presence of a rating is an important indicator, since points with only a higher rating will be added to my model. This will give our model a more accurate result, as it will give us an understanding that the point is more successful and that the analogue of such points should be opened in Iowa.

Chain	Av. Z score
Biggby	0.794
Jack in the box	0.537
Cracker-Barral	0.284
Smoothie King	0.193
Jamba Juice	0.126
Au bon pain	-0.033
Gloria Jeans	-0.120
Fuddruckers	-0.130
Peet's coffee & tea store	-0.381
In-N-Out Burger	-0.518
Pretzelmaker	-0.751

Table 10. Average Z scores for every chain

The results of the selection of the restaurant chain. After analyzing Z-scores, I determined that the most optimal network for my analysis is Biggby. This network has a fairly large number of points in the network, 96% of these points are in the Midwest. This network has no points in Iowa, which allows for a cleaner experiment. Also, it has reviews in Yelp for each point.

Case Study: Biggby Coffee. Biggby Coffee is a coffee chain which was founded in Michigan by Bob Fish and Mary Roszel in 1995. In the end of 1990s, they already had 3 locations and decided to develop their business as franchisee. Today Biggby coffee has 250 points in such states as Michigan, Florida, Illinois, Kentucky, Ohio, South Carolina, Texas, Wisconsin, and Indiana (“BIGGBY® COFFEE Locations” 2020).

They position themselves as a company with high-quality coffee. Coffee is selected by specialists. Coffee houses also offer a variety of methods of making coffee, which expands the audience of coffee. In addition to coffee, Biggby offers its customers a variety of muffins, cookies and yoghurts as an additional product ("BIGGBY COFFEE Franchise Review on Top Franchise Opportunity Blog" 2011). One of the advantages of Biggby is that they managed to reduce the cost of opening one point by almost 40%. Until 2011, the franchisee spent on opening a cafe amounted to more than $ 290,000. However, after the company revised the economics of using the retail space, the cost of opening one point fell to $ 210,000. This cost reduction also occurred due to the simplification of work with contractors, the search and delivery of materials, in other words, the model of supply and use of resources was revised ("BIGGBY COFFEE CUTS FRANCHISE START-UP COSTS 40%" 2011).

Today Biggby Coffee develops a new mini location franchisees model which is called B Cubed. The main idea is that the drive-thru/walk-up units are modular, taking up only 350 square feet, and can be installed in just 48 hours ("25 Best Coffee Franchises of 2020 (UPDATED RANKINGS)" 2017).

On the interview to "Shopping Center Buisness" magazine Biggby founder Mike McFall said that they don't pay too much attention to demographics because of their extraordinary success in Middle America.

He mentioned that they have successes in low income areas where there is high population density of working-class or middle class. For Biggby, McFall says, “anything north of $30,000 per year is good” (Scott Reid 2014).

After conducting a study of the website, the application and the concept of the cafe itself, I came to the conclusion that Biggby works for clients aged 35 and younger. Biggby has a huge number of flavors of drinks. The main ones are presented in the cafe, but most of them are in applications for smartphones. Also, their loyalty program is also in the application. This means t that they are aimed at a generation that will be more likely to download the application to their smartphones. In addition, the design of the cafe is dominated by bright colors and fairly loud music. This design also attracts people aged 20 to 30 years (Eric Decker 2016).

3.6 Café Suitability Factors

The factors that determine the location of the cafe are located in Table 1. This table shows the factors that were used by the researchers in drawing up the model of suitability for cafes and restaurants. Some of the factors (such as: pedestrian volume, convenience of garbage disposal, signage, public transportation (frequency per day) are only suitable for research in a fairly small area. When studying previous works, I entered the factors that the author indicated and put them in the table. Later I

counted the number of mentions of the factor in the articles and compiled a rating of the most popular factors among these works.

Study	Year	Income per person	Number of competitors	Population	Parking capacity	Traffic	Outdoor advertisement (m^2)	Visibility	Distance to public facilities	Transportation cost	Size of commercial area	Number of crimes
Chen et al.	2018		+		+		+		+			+
Yildiz and Tuysuz	2018	+	+	+								
Yang et al	2017	+		+								
Lin and Zu	2013	+	+	+		+						
Linder G. Ringo	2009	+			+	+	+	+				
Park and Khan	2006	+	+	+	+	+	+	+				
Tzeng et al.	2002		+		+				+	+	+	
Cella	1968	+		+		+		+		+		
Number of references		6	5	5	4	4	3	3	2	2	1	1

Table 11. Factors determining the location of the cafe according to scientific research

Most of the works were based on interviewing business owners and then compiling lists of the most topical critters when choosing a location for a business. To determine the most important criterion, they took a number of mentions of the criteria, and then determined the standard deviation from the mean. Thus, they got the value of how close the criterion is to the average value of mention (Park and Khan 2006). Yıldız Tüysüz did a similar study, but the survey was based on selected factors. Experts made a rating of factors for a retail store. Then the points were recalculated based on the standard deviation, and then a scale of importance of factors was built (Yıldız and Tüysüz 2018).

On the other hand, Lin and Zu also took the views of people adopting decisions to determine the coffee coefficient of importance for each criterion. Then the factors and coefficients were placed in the Hoff model, to determine the attractiveness of the territory for opening a café (Lin and Zu 2013).

Thus, based on the articles on the analysis of location, we can say that the most important factors are next: income per capita, number of competitors, population in the city, parking capacity and traffic. In our case, the factors "parking capacity" and “traffic” are important, since most of the population of Iowa travel by car, and public transport is poorly developed and does not have much popularity among the population.

When studying potential locations for the Dun Bros coffee shop chain, Ringo noted the importance of cafe visibility. This increases the likelihood that a person will go in and make a purchase (Linder G. Ringo 2009). Chen also motioned the importance of visibility and also outdoor advertisement. He added that due to the rules for installing outdoor advertising, which are different in many cities, competition is also changing (Chen et al. 2018). Thus, this parameter varies from city to city and will not be involved in my model.

Other factors such as: visibility, distance to public facilities, transportation costs, size of commercial area are less important based on previous articles, but also influence the choice of place. Unfortunately, it is not possible to get such data in the scale of 4 states. All these factors will not be included in my model.

In work on finding a place to open a restaurant in Taipei, Tzeng said that the cost of transportation and the availability of public space around plays an important role (Tzeng et al. 2002). He argues that it is necessary that the restaurant is in an area where a person can easily get without a car. In our case, the Midwest does not have a highly developed public transport infrastructure and almost all potential customers move in their own cars. Size of the commercial area factor was mentioned only once by Tzeng.

The number of crimes was also noted in only one source (Chen et al. 2018)but I consider this a rather important factor, which can also determine the potential position of the cafe. Information on the number of crimes is publicly available, so I will include it in my model.

Food franchise site selection variables

To have a full picture of criteria which can be used for opening a cafe, I have explored information of chain cafes and restaurants business development programs. Some of them have provided access to the franchise purchases sections. In this section it is possible to find the necessary parameters for the location where the franchise can start.

Company	Population	Traffic	Parking	Location	Visibility	Drive time	Income per person
Olive Garden	100000	20000	140	Freestanding			
Jack in the Box	20,000	25000	30	Signalized corner, out parcels, freeway locations	high visibility		
Panda Express	65000	45000	30	Heavy Retail, Employment, Shopping Centers; Entertainment		5-10	
Denny's	60000	25000	30	Theaters/restaurants		10	
Starbucks			20	Signalized corners with multiple access points			
Krispy Kreme	50000	25000		Outdoor patio seating in appropriate markets			60000
la boulangerie	50000	25000	20	Corner location, End Cap, Freestanding building			60000
Breakfast by Salt's Cure		Daytime and Evening traffic generators (theaters/restaurants/fitness)	20	Corner location, End Cap, Freestanding building Urban or Suburban storefronts, shopping centers, mixed use			80000

Table 12. Factors determining cafe locations according to franchise parameter

Thus, we see that for a business, the main parameters are population, traffic, parking capacity. Also, some of them have such parameters as: visibility, drive time and income per person. If we talk about the preferred location, then everyone will have it different, depending on the main customer base. However, this factor cannot be included in my model, since the Biggby coffee has its own priority locations and they will be indicated in the chapter on the analysis of the Biggby business model.

Biggby business model and site criteria

In addition to determining the necessary variables from an analysis of the past studies, as well as franchised models, it is necessary to understand how Biggby chooses its location. In an interview with Shopping Center Business Magazine, Mike McFall said that it takes about 18 months from the start of the search process to the opening. According to McFall, Biggby's business model, as a rule, does not imply the opening of coffee houses in detached buildings. The best filling for their coffee houses, Biggby considers the location at the ends of the linear shopping buildings. The size of the room must be 1,500 square feet, plus or minus 20 percent (1,200 to 1,800 square feet)

According to McFall the Biggby Coffee looks for new places by using next steps:

1. Count transient populations, how many people are driving by a location in a day.

2. Analysis of population density, or how many people live one mile from the store.
3. A daily population count showing the number of people who come to work within a radius of one mile.
4. Good retail add-ons, including those in the general area that drive traffic (Scott Reid 2014).

Final list of variables for modelling

Based on research on the study of the choice of location, as well as information from franchises, it is possible to make a final set of variables, which I will use in my model to search for a location using machine learning.

Socio-economic block. This block includes data such as: total population, median disposable income, median household income, population density, percent of working population, percent of American Indigenous population, percent of Asian population, percent of Black population, percent of Hispanic population, percent of White population. The data was taken from the ESRI database and generated on the scale of the census tract. Specific data was taken in order to comply with the business model for finding new locations in Biggby

Eating places block. This block includes data such as: restaurants in one mile radius, cafes in one mile radius, fast food in one mile radius. Data

were collected from the Open Street Maps. The relevance of the data comes from the availability of data on this resource at the time of the study. The data were divided into three groups so as to better understand which type of met catering is most important

Road data block. This data block includes data such as: number of parking lots, and the number of intersections in one mile radius. The relevance of the data comes from the availability of data on this resource at the time of the study. The use of road also comes from the Biggby's business model.

Retail add-ons block. This data block includes data about retail which is located in one mile radius: services car shops, children shops, clothes shops, convenience stores, department stores, electronics stores, entertainment stores, furniture store, gas stations, grocery stores, pharmacy and health centers, goods for home stores, malls, stores with "No_info", stores with category "Other", pet stores, sports stores, supermarkets, Tobacco shops, vacant stores, variety stores, wholesale stores. The data was also divided into groups in order to better understand which type of retail is more influential in the appearance of a new Biggby point. All model inputs in figure format can be found in appendix.

Variable	Description	Source
Total population	Total population by census tract	ESRI 2019
Median disposable income	Median disposable income by census tract	ESRI 2019
Median household income	Median household income by census tract	ESRI 2019
Population density	Total population divided by census tract area	Calculated by using ESRI 2019 and data from Census Bureau
Percent of working population	Total number of workers divided by total population n census tract	ESRI 2019
Percent of American indigenous population	Total number of American indigenous population divided by total population in census tract	ESRI 2019
Percent of Asian population	Total number of Asian population divided by total population in census tract	ESRI 2019
Percent of black population	Total number of black population divided by total population in census tract	ESRI 2019
Percent of Hispanic population	Total number of Hispanic population divided by total population in census tract	ESRI 2019
Percent of white population	Total number of white population divided by total population in census tract	ESRI 2019
Number of parking lots	Sum of parking lots in a radius of one mile around the cell	Calculated by using OSM data
Restaurants in one mile radius		
Cafes in one mile radius		
Fast food in one mile radius		
Services in one mile radius		
Car shops in one mile radius		
Children shops in one mile radius		
Clothes shops in one mile radius		
Convenience stores in one mile radius		
Department stores in one mile radius		
Electronics stores in one mile radius		
Entertainment stores in one mile radius		

Furniture store in one mile radius	Sum in a radius of one mile around the cell	Data from Open Street Maps
Gas stations in in one mile radius		
Grocery stores in one mile radius		
Pharmacy and health centers in one mile radius		
Goods for home stores in one mile radius		
Malls in one mile radius		
Stores with “No_info” in one mile radius		
Stores with category “Other” in one mile radius		
Pet stores in one mile radius		
Sports stores in one mile radius		
Supermarkets in one mile radius		
Tabaco shops in one mile radius		
Vacant stores in one mile radius		
Variety stores in one mile radius		
Wholesale stores in one mile radius		
Number of intersections in one mile radius		

Table 13. Final variables determining cafe locations

CHAPTER 4

RESULTS

The results that will be presented in this chapter are divided into several blocks:

The first block describes the results of the preliminary analysis data from all states (Michogam, Indiala, Ohio and Iowa) and high resolution using Maxent algorithms. The second block presents the calculation of collinearity coefficients and the selection of variable variables that do not correlate with each other.

The third block describes the result of analyzing the suitability of a location using a logistic model in the Maxent program and comparing it with the results of data analysis using models in Python. This analysis was based on data from only two states: Iowa and Michigan.

4.1 Model 1. Maxent for All States

The first result was obtained using logistic regression in the Maxent program. This model was compiled from data for four states: Iowa, Michigan, Ohio and Indiana. This model was calculated on the basis that the variables were taken on the basis that one cell is 100 per 100 meters or one section in a block. Thus, the area of all four states was divided into cells of 100 per 100 meters. As variables, all variables that were listed in the chapter methods were used. Also, the data set about the locations of

the Bigbby location was divided into 2 parts: data for training the model (80% of the points) and data for testing the model (20% of the points).

The percentage of a random test, which is 20%, allowed the program to do a simple statistical analysis. Most of the analysis was built on using a threshold to make binary prediction. Suitable conditions are presented above the threshold, and unsuitable conditions are presented below. Figure 39 shows how testing and training are missing out and the projected area varies depending on the choice of the cumulative threshold.

Here we see that the skip on the test samples is higher than the predicted skip frequency for the skip coefficient for test data taken from the Maxent distribution itself. The coefficient for the forecast of omissions is depicted by a straight line to determine the total output format (Phillips and Research 2017).

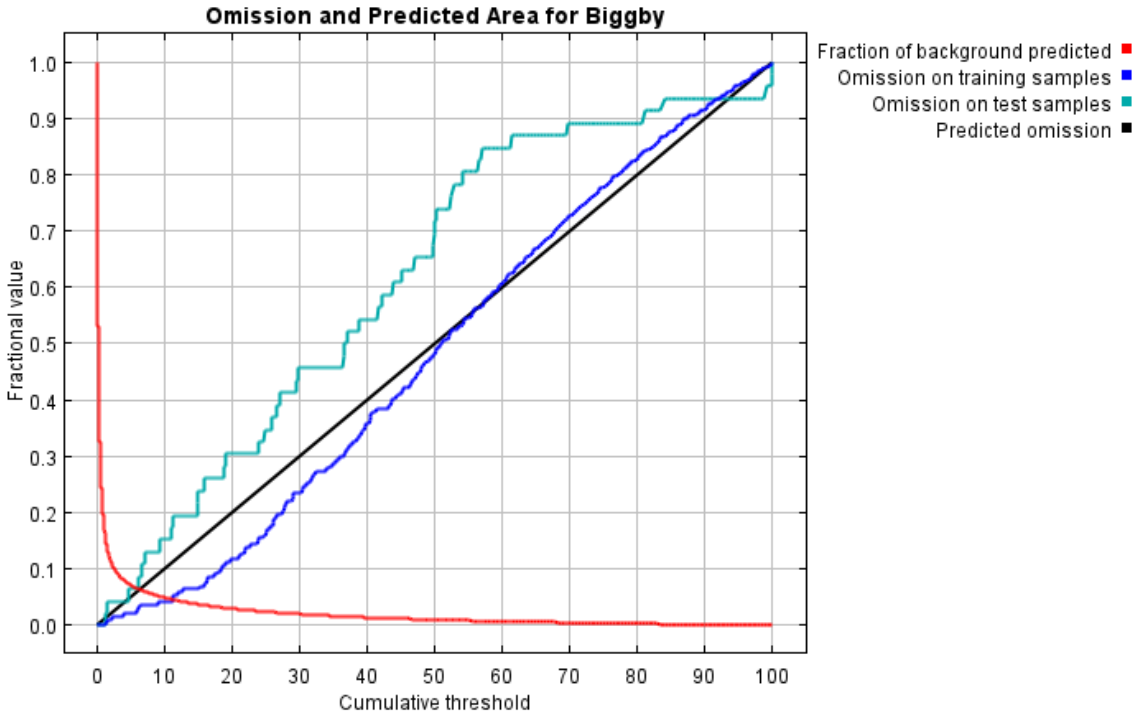


Figure 1. Omission and Predicted area for Biggby

In some situations, the skip line of the test lies well below the predicted skip line: the common reason is that the test and training data are not independent, for example, if they are obtained from the same spatially auto correlated presence data. In our situation, the skip line of the test is completely above the forecast line, so our data are essentially independent. But at the same time it shows us that based on test data, you can see that the algorithm assigns more data to a negative class than to a positive one.

Another important indicator for checking regression results is Area under the Receiver Operating Characteristics (AUROC). The AUC - ROC curve is an accuracy measurement for classifying data. ROC is the probability curve, and AUC is the degree or measure of separability shows how well the model is able to distinguish between classes. The higher the AUC, the better the model when predicting 0 s 0 and 1 s 1 s (Phillips and Dudık 2007). In our case AUC for training data is 0.98 and for testing data is 0.97. That means that accuracy for our model is 98% for training data and 97% for testing data. All 98% of classification for point which was made by Maxent logistic regression was right.

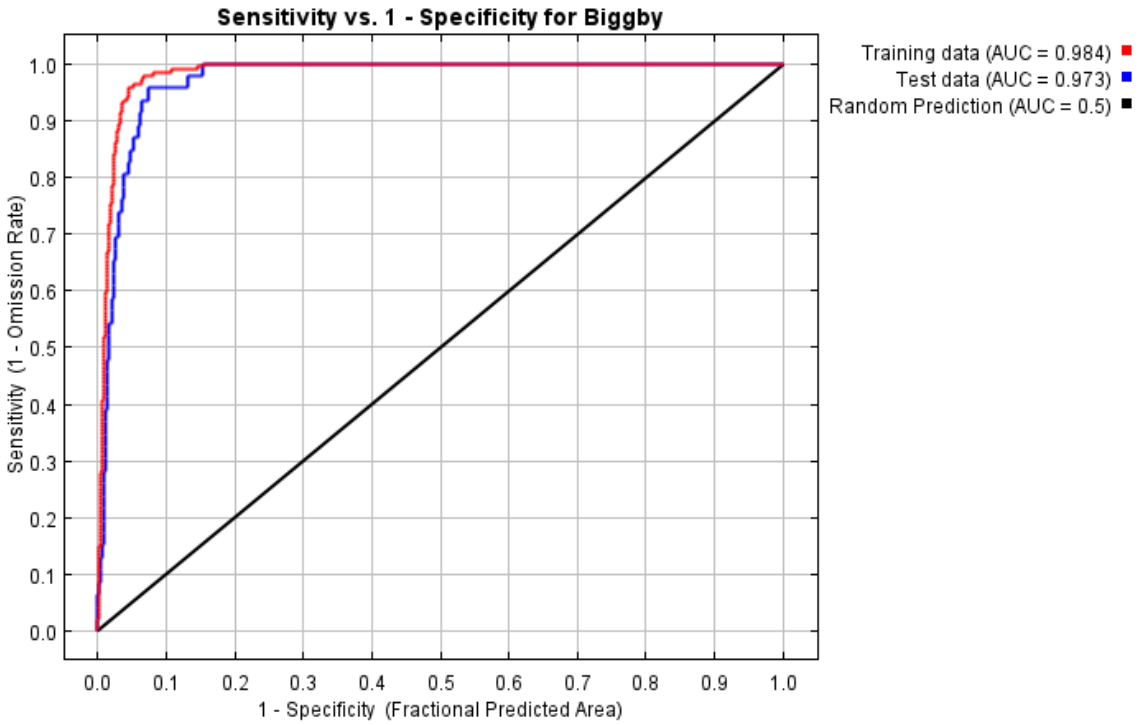


Figure 2. AUC curve for Model 1

Another important indicator is the determination of the contribution of each of the variables to the classification of logistic regression. Contribution values are determined only heuristically. The result depends on the path that Maxent uses to solve the classification problem. Another algorithm may lead to different results. Also, due to the possibility of having variables that correlate with each other, the results may also not be the most accurate. According to Phillips (2006), the percent contribution of each variable is calculated as follows: "While the Maxent model is being trained, it keeps track of which environmental variables are contributing to fitting the model. Each step of the Maxent algorithm increases the gain of the model by modifying the coefficient for a single feature; the program assigns the increase in the gain to the environmental variable(s) that the feature depends on. Converting to percentages at the end of the training process, we get the percent contribution."

The right column in the table shows the second indicator of the contribution variable, which is called the importance of the permutation. This measure depends only on the final model of Maxent, and not on the path along which it was obtained. The contribution for each variable is determined by randomly rearranging the values of this variable between training points (both presence and background) and measuring the resulting decrease in training AUC. A large decrease indicates that the

model is highly dependent on this variable. Values are normalized to give percentages (Phillips and Research 2017).

Variable	Percent contribution	Permutation importance
Fast Food	38.5	0.1
Intersections	24.9	50.2
Restaurant	15.1	0
Parking	6.3	5.6
Population density	5.8	8.6
% of american ind. pop	1.3	7.7
Supermarkets	1	0.6
% of_hispanic_pop	1	5
% of_white_pop	1	0.1
Department stores	0.8	0.6

Table 14. Analysis of variable contributions

The table 14 shows the top 10 most important variables for which the Maxent algorithm received the highest data for contribution percent. Thus, we see that in our model the most important variable is the number of fast food points within a radius of one mile from the cell. However, in permutations, the role of this variable is not significant. Another important variable is the number of intersections within a radius of one mile from the cell. At the same time, we see that from the rearrangement of the values of this variable, the value of AUC is experiencing a strong change. Thus, it can be understood that the location of the points near the crossings is one of the most important parameters. Other important parameters are: the number of restaurants, as well as the number of parking spaces within a

radius of one mile from the cage. Population density is also important, however, according to the result of the algorithm, the percentage of working people in the tract has an indicator 0.7 for contribution percent, but it has 2.1 for permutation importance. A rather surprising indicator is the significance of the percentage of the indigenous population. At first glance, it may seem that this variable correlates with another, but the Pearson test did not show the dependence (the coefficient did not rise above 0.6). The remaining variables do not look unusual, but are also significant for the model.

4.2 Visual Assessment of Model 1 Results

One of the results of data analysis using Maxent is a probability map. First of all, we will consider a general map with the results for all four states. We see that basically the highest probability of the location of a new point of Biggby Cafe is in urban areas. Of course, the greatest amount of the highest probability is in the vicinity of the largest cities, such as: Detroit, Indianapolis, Columbus and Des Moines. Also, the probability increases in small cities. Perhaps this is due to the fact that even in small towns the number of crossings increases significantly and therefore even such places are identified as suitable with a high probability. In large and medium-sized cities, locations are mostly confined to shopping areas, where there is a large concentration of various services and shops. Also, business centers and downtowns, where there is a large crowd of people,

are highly suitable. At the same time, it is possible to observe a decrease in probability in the vicinity of the business center. Such a picture, for example, can be observed in Detroit. Indianapolis, Des Moines and Columbus have the same situation. The likelihood increases in the city center and further decreases in residential areas located closer to the center.

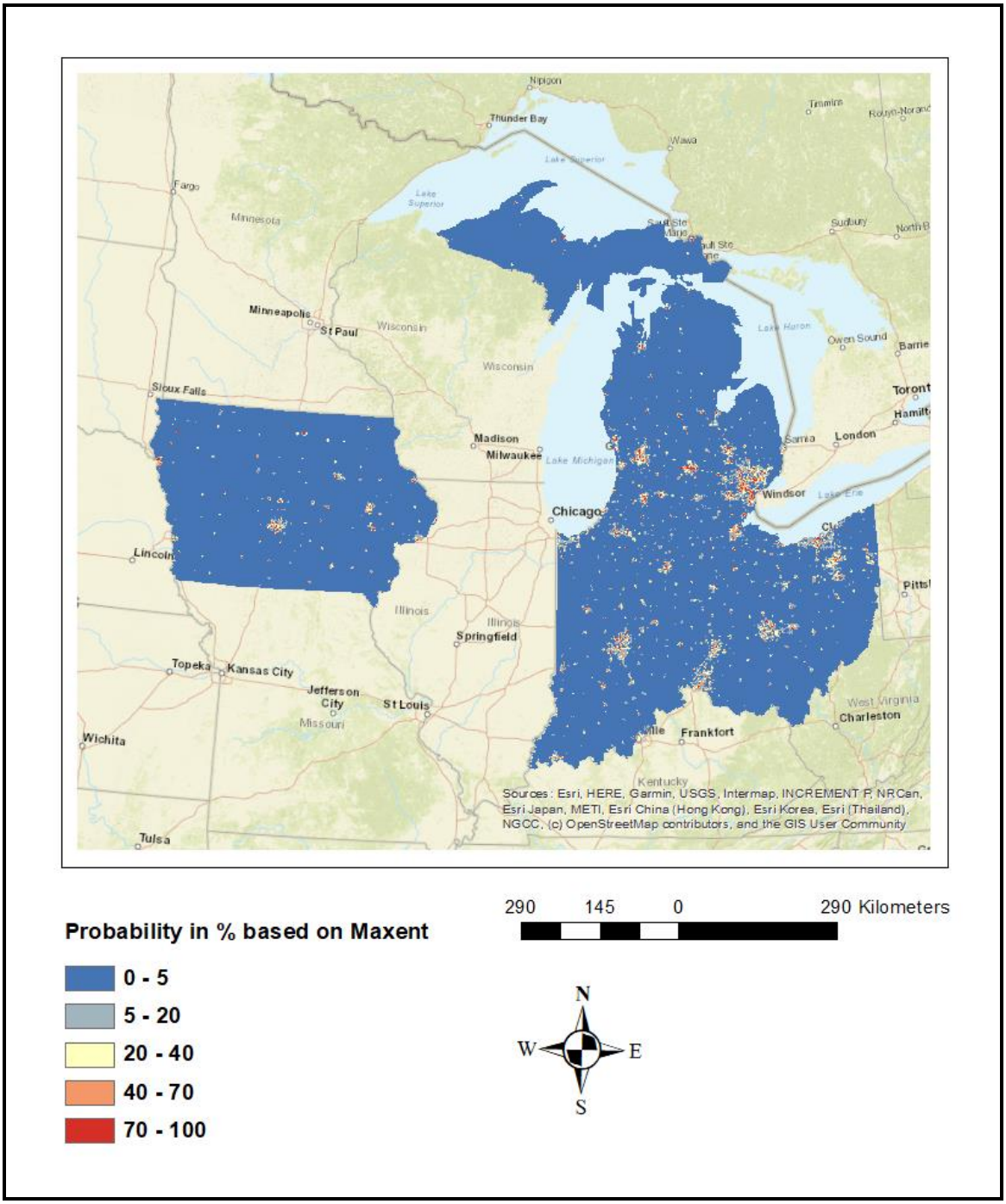


Figure 3. Maxent model for Biggby cafes for Iowa, Michigan, Ohio and Indiana

For Des Moines (large city) Maxent gave a fairly detailed forecast. The probability was not evenly distributed throughout the city. The downtown area, the zone in the north in the Marquisville region, was most likely. In this area there are quite a lot of outlets. Also, the shopping area in Jordan Park received high suitability. Also, the commercial zone in Koh'l West Des Moines and Hickman Road received high suitability. If we compare the results with the zoning map of the city, then it turns out that the Maxent practically predicted all the zones in which it is possible to open a catering place.

If we consider the result for Cedar Falls (medium-sized city), it is possible to see that the model made its prediction accurately enough. The algorithm assigned the University Avenue, Main Street and the shopping area in East Viking Road, where Walmart and other large stores are located, with high probability. In this area there is a large number of cafes, restaurants and fast food. Also in this area are supermarkets. There are restaurants and cafes in the Main Street area. Thus, if we talk about the strategic analysis of locations on a statewide basis, the algorithm predicts quite accurately not only cities or macro zones, but also areas within small cities.

For Waverly (small city), the Maxent quite clearly identified the downtown areas, but at the same time the area near Warburg College,

where a large number of cafes are also concentrated. A trade area in the south of Waverly was also allocated.

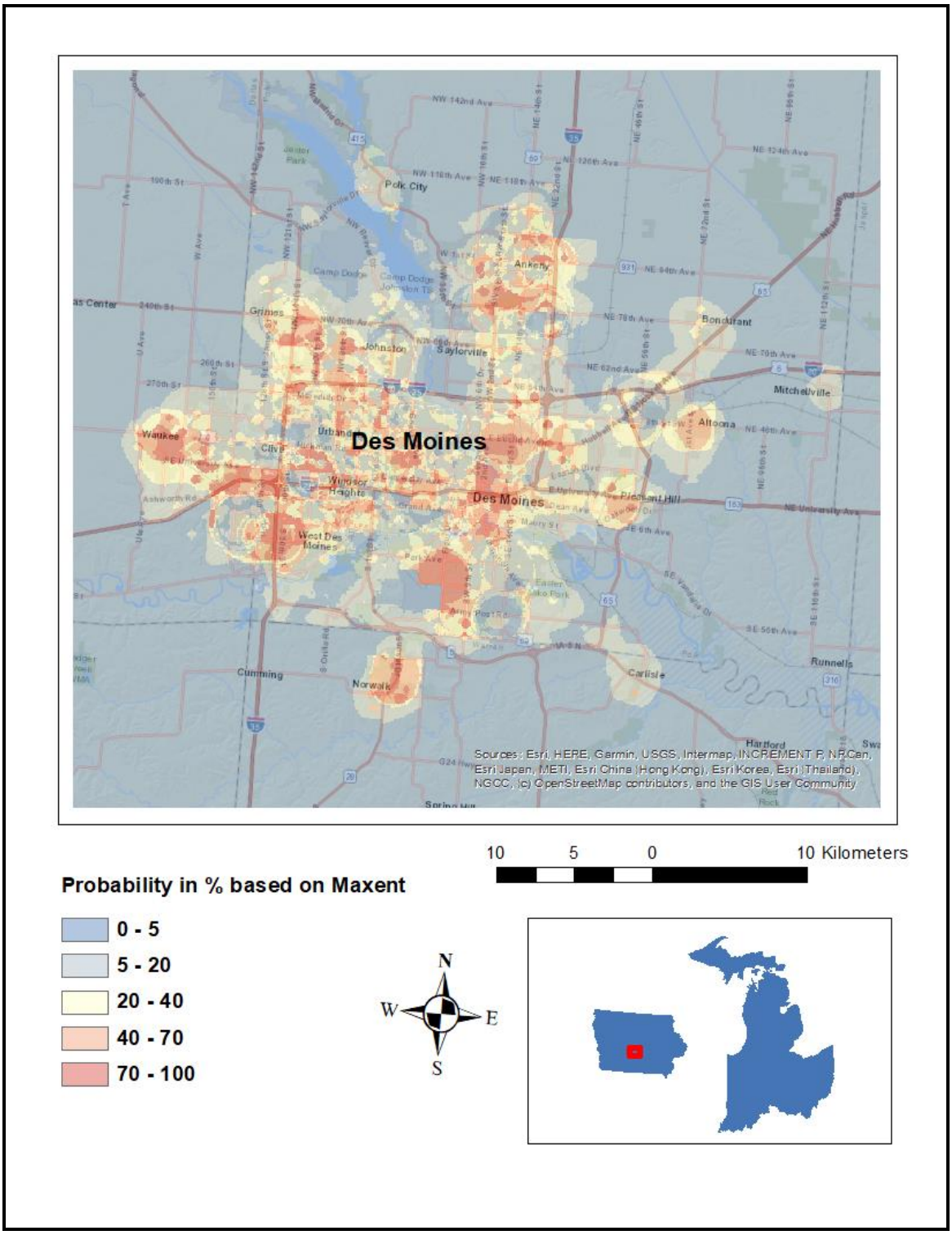


Figure 4. Maxent results. Zoom to Des Moines

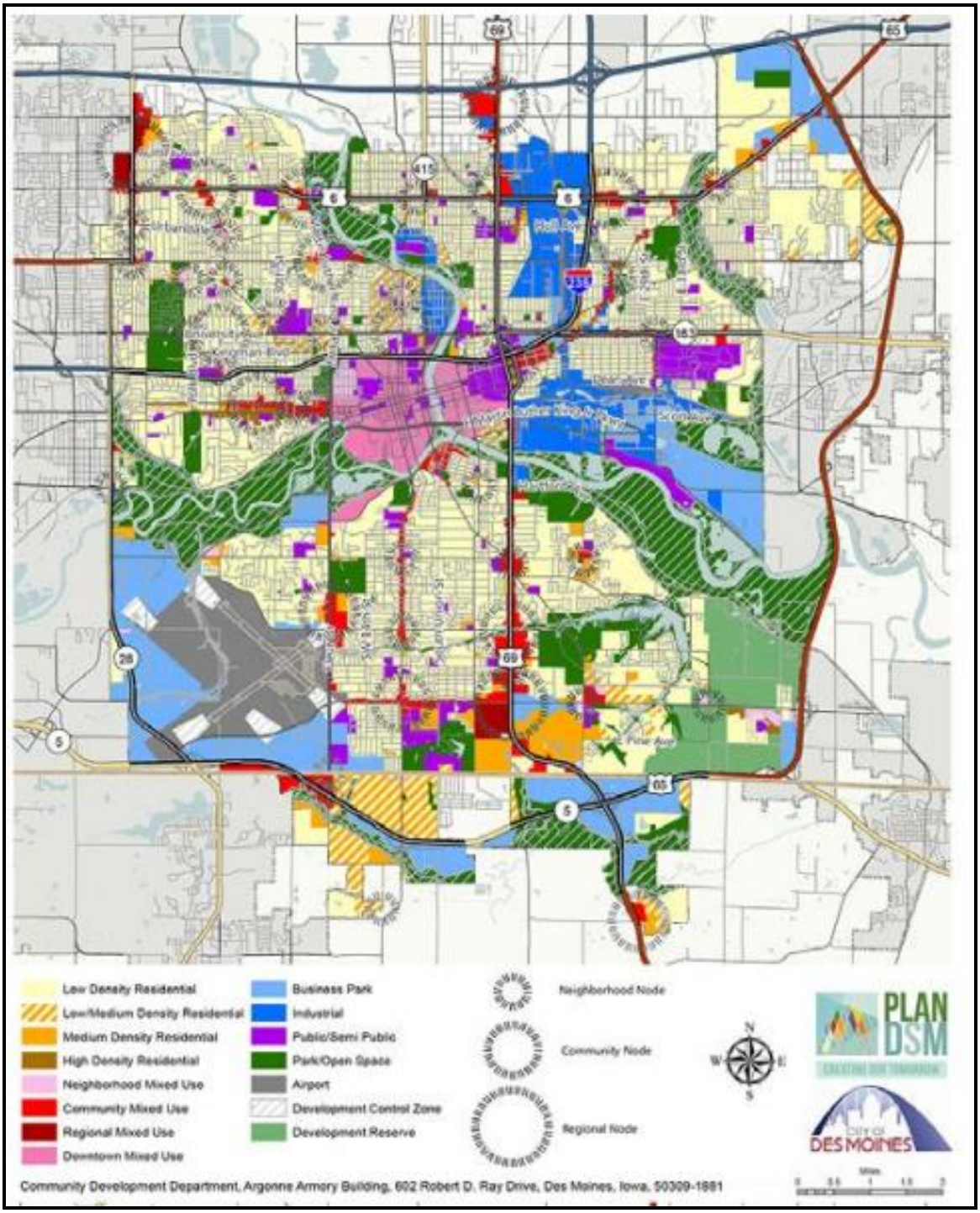


Figure 5. City of Des Moines comprehensive Plan

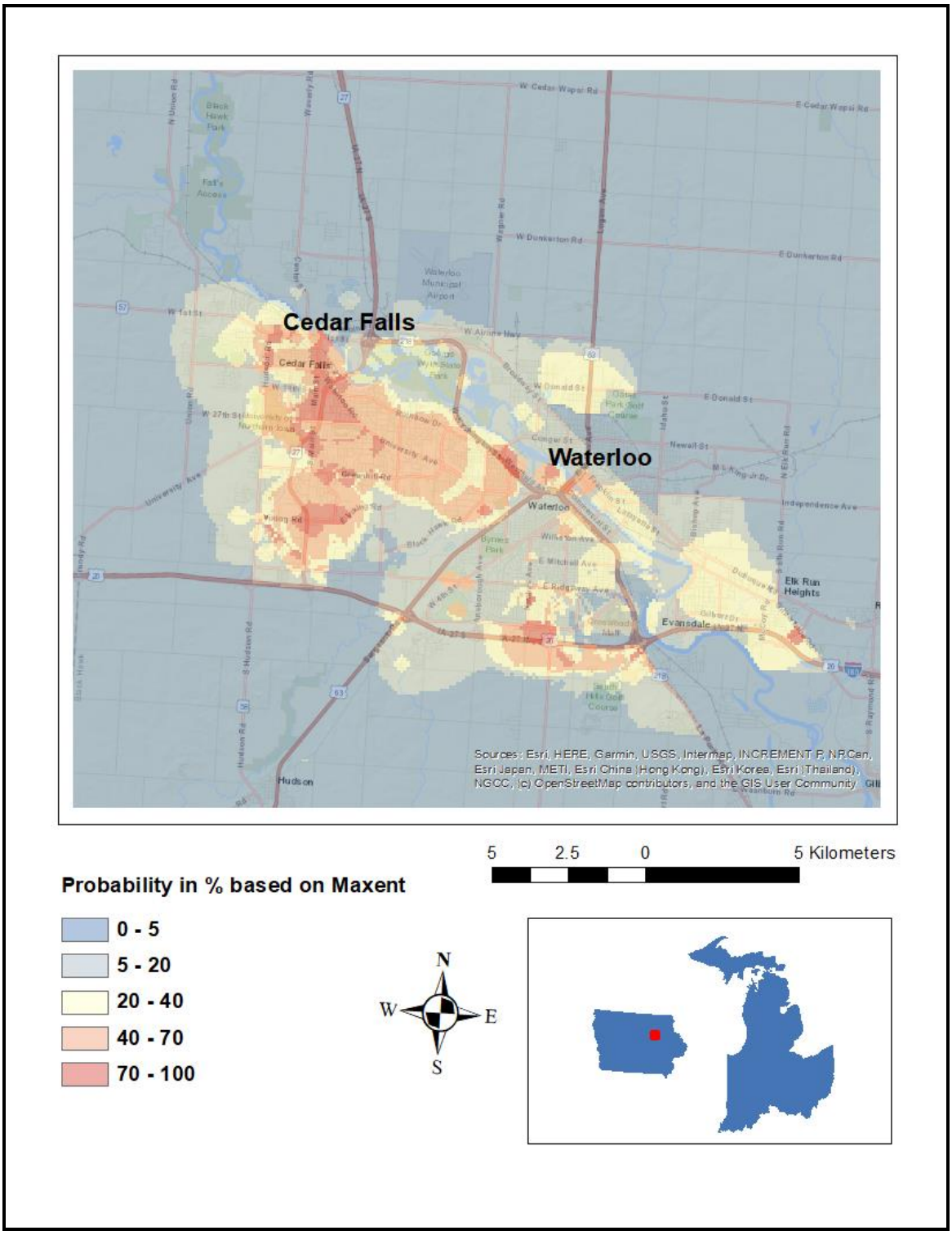


Figure 6. Maxent results. Zoom to Cedar Falls

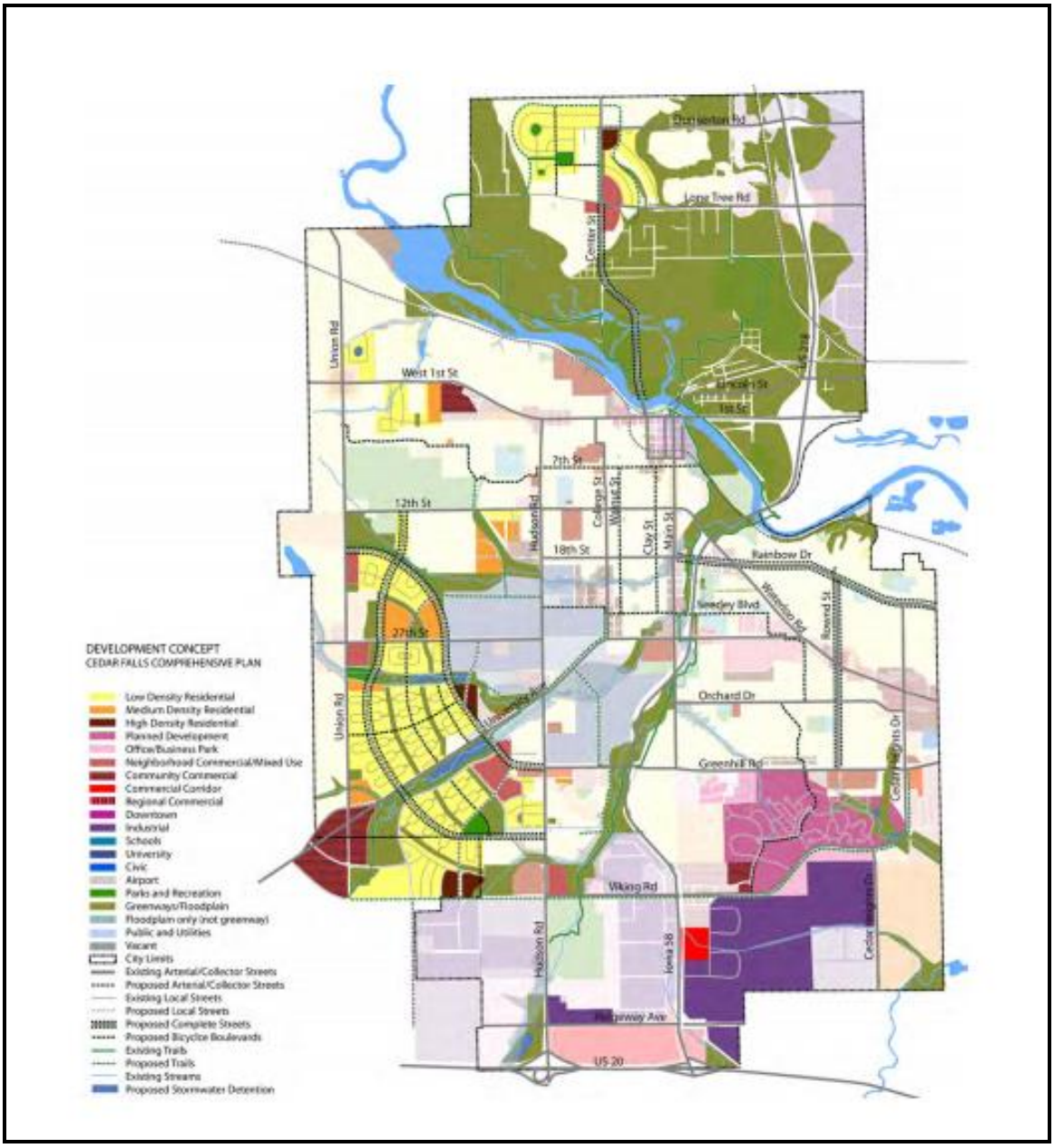


Figure 7. Comprehensive Plan for the City of Cedar Falls

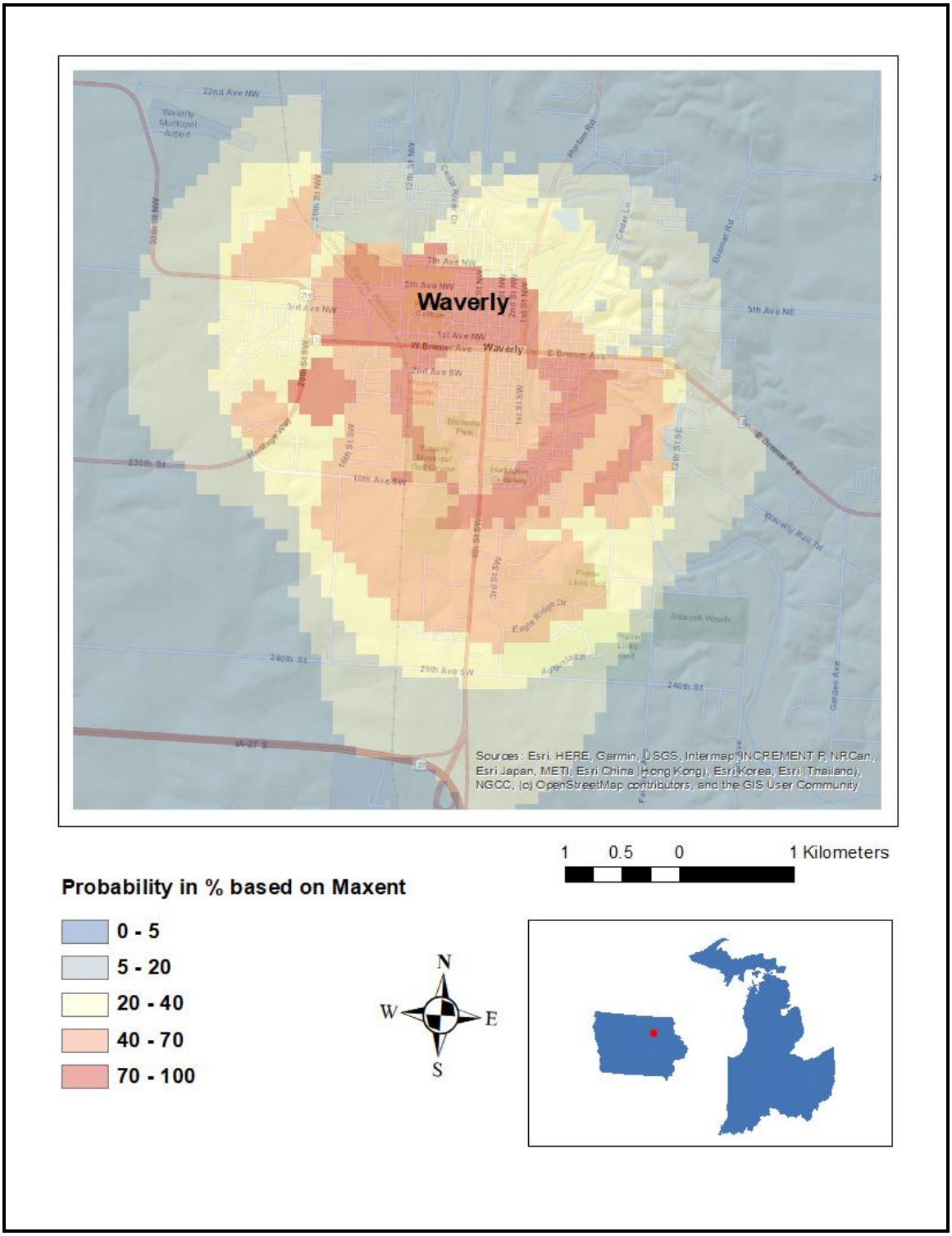


Figure 8. Maxent results. Zoom to Waverly

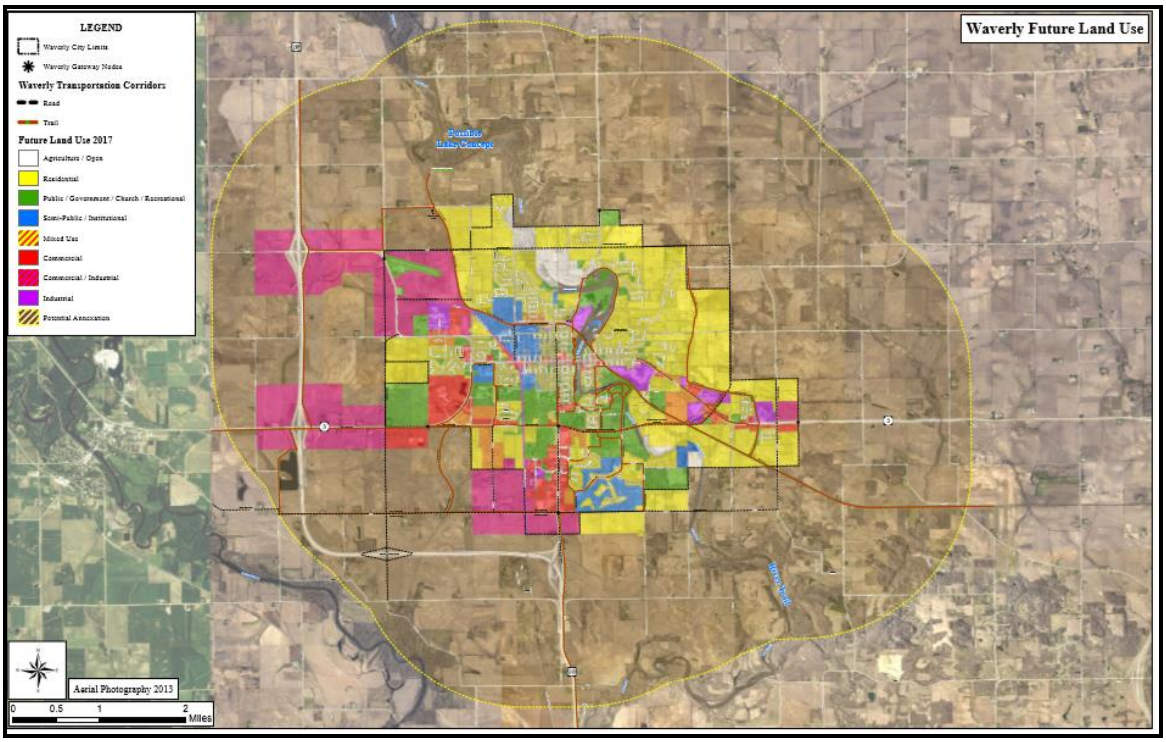


Figure 9. Comprehensive land use plan of Waverly

Adjusting the scale of analysis

Previous calculations were based on the assumption that all 4 states were divided into 100 cells by 100 meters. When translating this into a table format, it turned out that the number of rows in the table exceeds 50 million. The table contains 38 variables. All this created difficulties for calculations. Therefore, it was decided to increase the area of one cell to 300 by 300 meters, which equals the length of one residential block. It was also decided to abandon the use of Ohio and Indiana data, as more than 90% of Biggby cafe points are concentrated in Michigan.

4.3 Estimation of Correlation between Variables and Collinearity Analysis

The previous result was obtained without measuring how different variables affect each other. Inside the Maxent algorithm, a check is underway on how a particular variable affects the final result and how the AUC changes. However, to obtain a more accurate result, it is necessary to check how different variables correlate with each other, thereby changing the final forecast.

The test was carried out in two ways: the test for multicollinearity in R. and measuring the Spearman coefficient for all variables. Correlation was estimated using the Spearman coefficient in Python using the pingouin package. At the heart of this package are Pandas and Numpay. The package makes it easy enough to do statistical calculations (Vallat 2018). All of the following calculations and results were made for Iowa and Michigan with a cell size of 300 by 300 meters. Script of calculation Spearmen correlation can be found in appendix.

The formula for Spearman correlation is next:

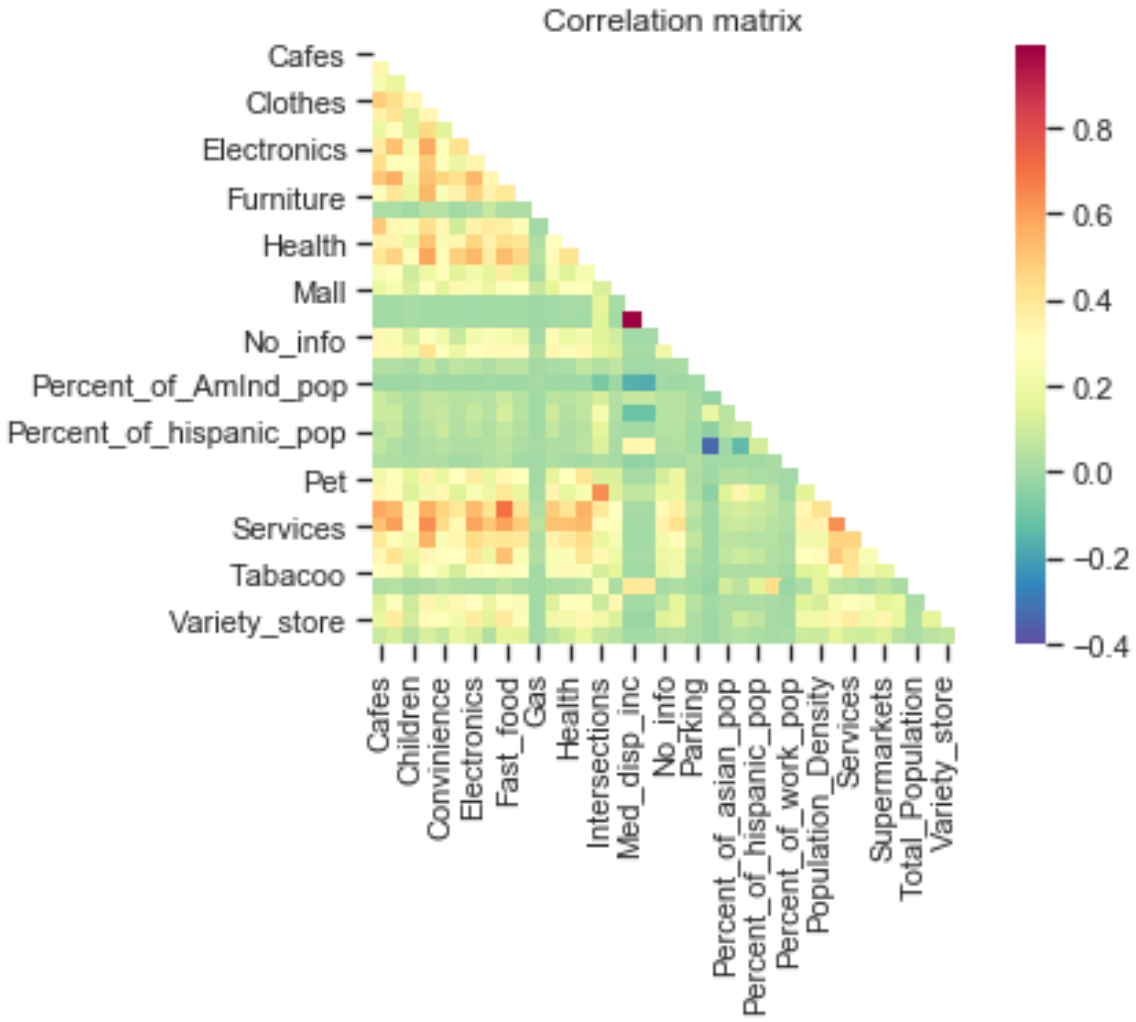
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Whrere $\sum d_i^2$: - sum of squared rank differences, n^2 - number of paired observations.

When using the coefficient of rank correlation conditionally assesses the tightness of the relationship between the signs, considering the values

of the coefficient equal to 0.3 or less, indicators of weak tightness of communication; values of more than 0.4, but less than 0.7 - indicators of moderate communication tightness, and values of 0.7 and more - indicators of high communication tightness (Polyakov 1971).

As a result, a matrix was obtained with coefficients for all variables that participate in the study.



Plot 1. Matrix of Spearman correlation index

Thus, we see that variables showing the number of restaurants within a radius of one mile around a pixel and a variable showing the number of places with different services, also within a radius of one mile, are highly correlated (index = 0.641). There is also a cross-correlation between the location of restaurants and fast food (index = 0.706). The correlation is traced between shops with the vehicle industry (sale of cars, spare parts,

car workshops) as well as companies that provide various services (index = 0.602). There is also a relationship between services and clothing stores (index = 0.645).

Another test was the collinearity diagnostics. Collinearity or excessive correlation between variables can make it difficult to find the optimal set of variables for a given model. The usual approach to selecting variables can lead to conflicting results that reduce the accuracy of the model.

It is difficult to resist the temptation to build an environmental model using all available information (i.e. all variables). The analytical constraints associated with collinearity require us to carefully consider the variables that we choose for modeling, and not the naive approach, when we blindly use all the information to understand the complexity.

A simple approach to determining collinearity among explanatory variables is to use dispersion inflation factors (VIF) ("Collinearity and Stepwise VIF Selection – R Is My Friend" 2013). VIF calculations are simple and straightforward; the higher the value, the higher the collinearity. The VIF for one explanatory variable is obtained using the r-squared regression value of this variable relative to all other explanatory variables:

$$VIFj = \frac{1}{1 - R_j^2}$$

Where the VIF for variable j is the reciprocal of the inverse of R^2 from the regression. A VIF is calculated for each explanatory variable and those with high values are removed. The definition of 'high' is somewhat arbitrary but

values in the range of 5-10 are commonly used ("Collinearity and Stepwise VIF Selection – R Is My Friend" 2013).

Functions for Medical Statistics Book (fmsb) package was used for creatinf VIF test for our variables. Script can be found in appendix.

As a result of the analysis, only 2 variables received a value greater than 5.

Variable	VIF
#Med_disp_inc	126.864417916284
#Med_hh_inc	125.926995050513

Table 15. Variables with the highest coefficient value VIF

It turns out that the VIF test identified only two variables as correlating with each other. This correlation is fairly obvious, as it is likely that there will be more people with the highest amount of disposable income in areas where there is more median household income.

Based on the two tests, we see that in our data set there are several variables that should be removed to obtain a more accurate result: median disposable income, restaurants and services. After these tests my variable list will contain next parameters:

Vatiable name	Calculation method
'Med_hh_inc'	Value without additional calculation
'Percent_of_AmInd_pop'	The ratio of the desired variable to the total number of people
'Percent_of_asian_pop'	Value without additional calculation
'Percent_of_black_pop'	The ratio of the desired variable to the total number of people Number of people divided by tract area Total population in tract
'Percent_of_hispanic_pop'	
'Percent_of_white_pop'	
'Percent_of_work_pop'	
'Population_Density'	
'Total_Population'	
'Parking'	Sum of parking lots in a radius of one mile around the cell
'Car shops'	Sum in a radius of one mile around the cell
'Children shops'	
'Clothes'	
'Convinience'	
'Department_store'	
'Electronics'	
'Entertaiment'	
'Furniture'	
'Gas station'	
'Grocery'	
'Health'	
'Home'	
'Mall'	
'No_info'	
'Other'	
'Pet'	
'Sports'	
'Supermarkets'	
'Tabacoo'	
'Vacant'	
'Variety_store'	
'Wholesale'	
‘Cafes'	
'Fast_food'	
‘Intersections’	

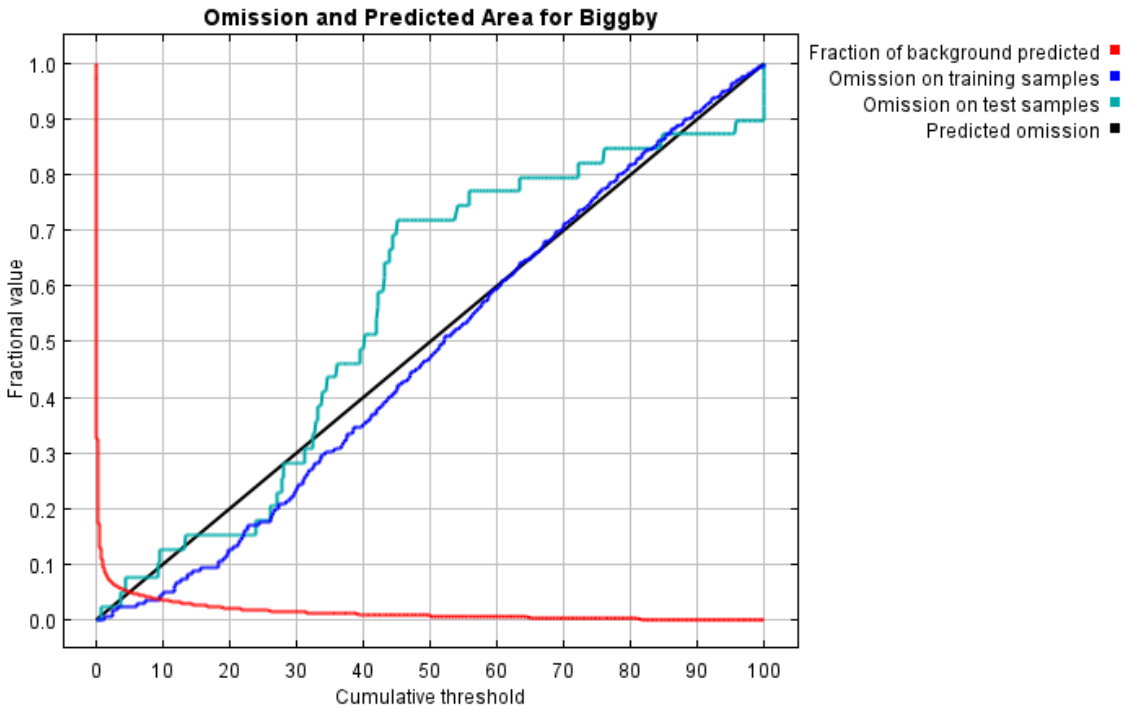
Table 16. Variable list after correlation test

4.3 Model 2. Maxent Model with Resolution of 300 x 300 Pixels

These results are based on data for the states of Iowa and Michigan. The area of one cell is 300 by 300 meters. For Maxent the existing location

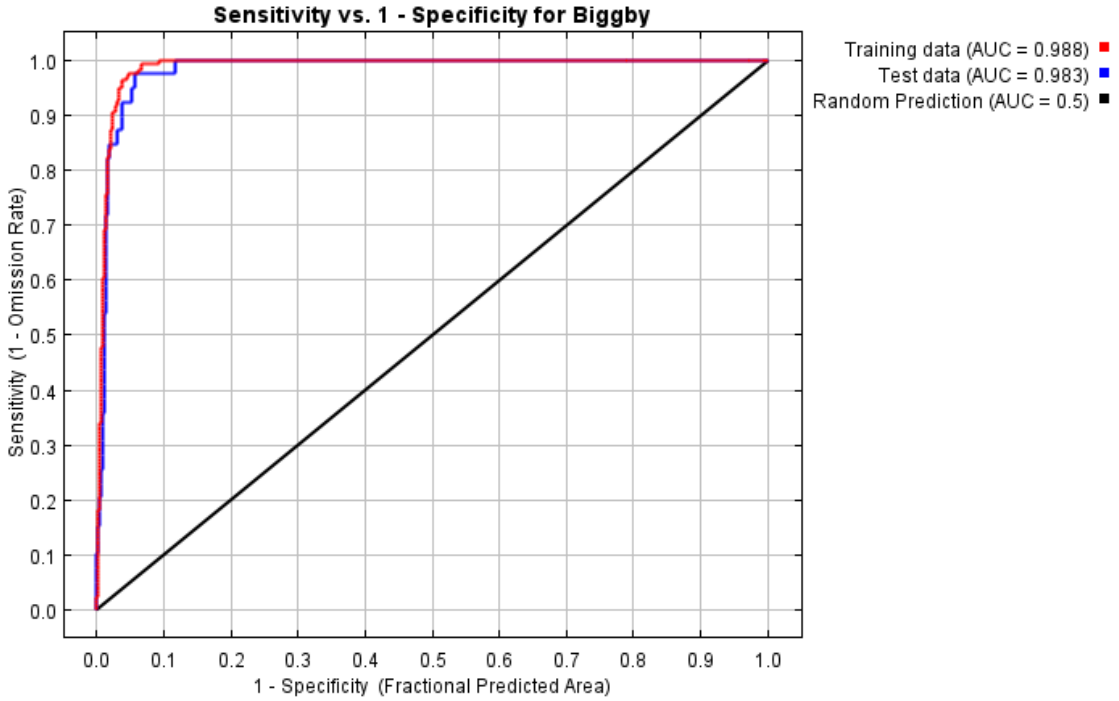
points of Biggby Cafe are divided in the proportion of 80% for training and 20% for testing the result

Analysis of omission /commission. Compared to the previous result, the test line is closer to the predictive omission. This means that this time the regression more accurately distinguished the results into two classes. However, in this case, one can also see that, as in Model 1, the regression characterizes most cells as a negative cluster, that is, as unsuitable for the appearance of points in a Biggby cafe. However, this is predictable, since most of the important variables are in cities that occupy a small area compared to the state. It is also seen that the regression assigned fewer values to 1, that is, the algorithm predicted fewer points with a probability of 100%.



Plot 2. Omission and predicted Area for Biggby

In this case AUC for training data is 0.99 and for testing data is 0.98. That means that accuracy for our model is 99% for training data and 98% for testing data. All 99% of classification for points which was made by Maxent logistic regression were right. This result is more accurate than Model 1.



Plot 3. Sensitivity vs 1 – Specificity for Biggby

Analysis of variable contributions.

In this case, the result of the contributions of the variable to the predictions of the model also changed. In Model 1 the most important variable was the number of fast food restaurants, intersections and restaurants within a radius of one mile of the cell. This time the population density variable has become the main more important than restaurants and interactions variables. But the intersections variable is still most important by the permutation parameter.

Variable	Percent contribution	Permutation importance
fast_food	33.6	2.2
population_dansity	31.7	24.4
intersections	27.5	50.9
med_hh_inc	1.2	0.7
supermarkets	0.7	1.6
cafes	0.7	1.3
percent_of_work_pop	0.6	0.9
percent_of_amind_pop	0.6	1.5
no_info	0.6	1.4
convinience	0.4	0.6

Table 17. Top 10 variable by importance in Maxent model

4.4 Visual Assessment of Model 2

Visually, it can be observed that in Model 2, the number of zones with a high probability has become less. Zones have become more aggregated. If the first result predicted that most small towns are also a good place to open cafes, now many small towns in Michigan are no longer significant. Iowa has not undergone major changes, but areas of high suitability have also become smaller. If we compare the percentage of cells with different probability for our two models, we will see that the percentage of high probability cells in the first Maxent result is 0.365% and for second is 0.384%. To check whether these differences are significant, I conducted a *z* test (a test for two proportions), which showed that the value of p is less than 0.05. This means that the differences are significant.

Second Maxent result			
Probability value	Percent	Probability value	Percent
0.7 - 1	0,365%	0.7 - 1	0,384%
0.4 - 0.7	0,968%	0.4 - 0.7	0,566%
0.2 - 0.4	1,417%	0.2 - 0.4	1,161%
0.05 - 0.2	3,606%	0.05 - 0.2	7,854%
0 - 0.05	93,643%	0 - 0.05	90,036%
Grand Total	100%	Grand Total	100%

Table 18. Proportion of cells for each class for first and second Maxent results

For Des Moines, Maxent made a similar forecast with the previous model. For Des Moines, Maxent made a similar forecast with the previous model. However, many areas with a high probability have become larger and are now merging. So the downtown area has become more suitable and now occupies almost the entire territory, extending to the western part. Also increased area in the area of Clive

In Cedar Falls we also see a decrease in the number of cells with high probability. There remains one main area on East Viking road, next to the trade cluster where Walmart is located. For Waverly, the precision of the forecast has greatly decreased and now the zone with a high probability covers the entire city, which reduces the precision of the model.

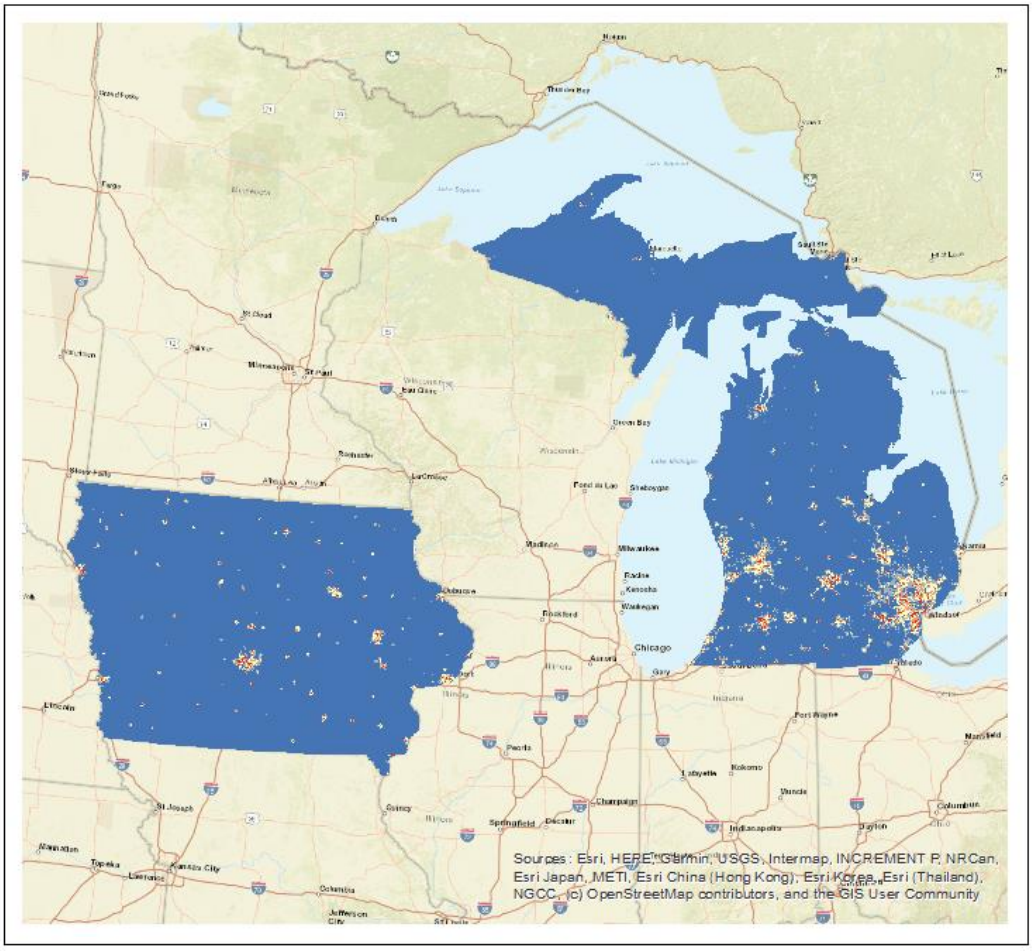


Figure 10. Maxent results for Iowa and Michigan after correlation test

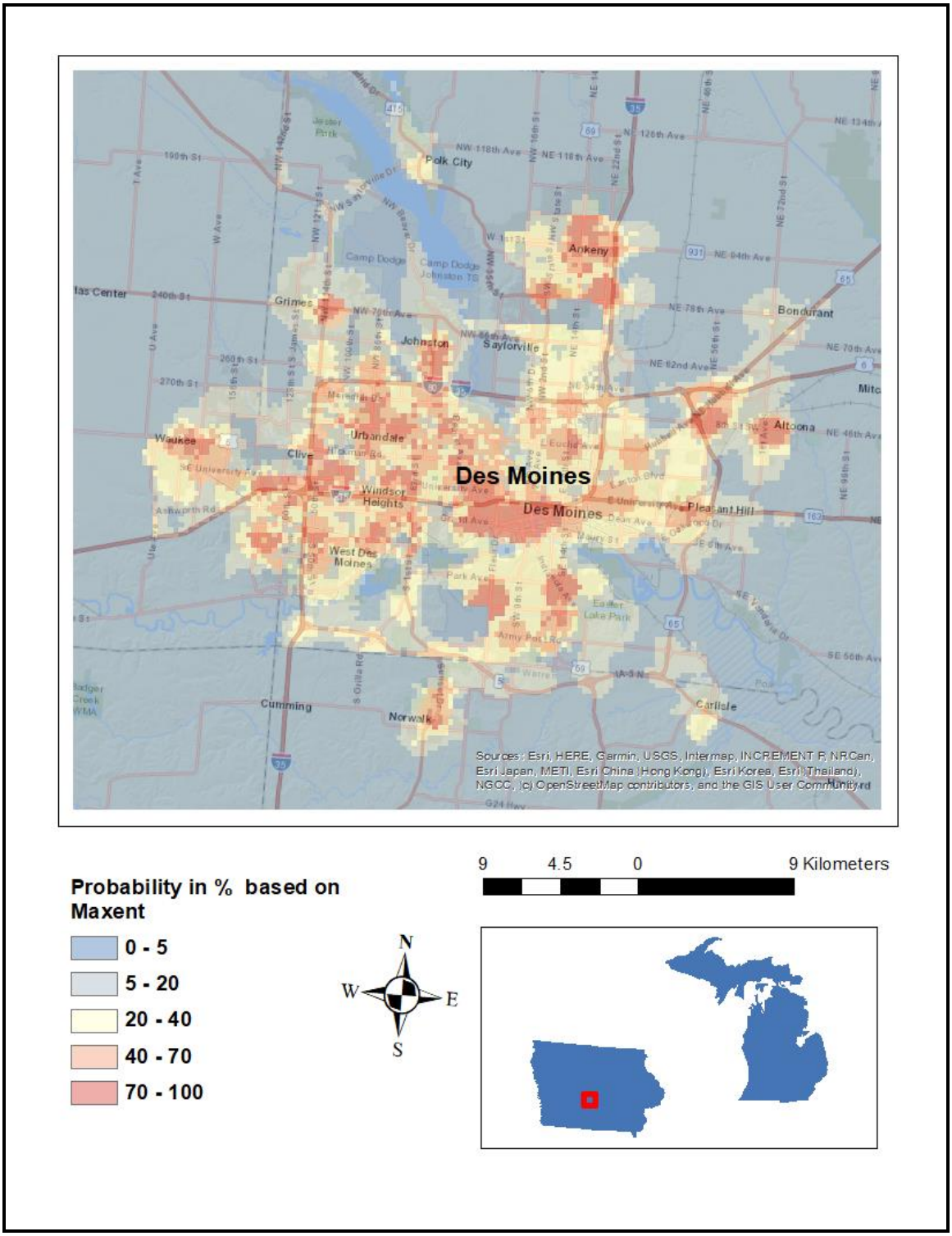


Figure 11. Maxent results after correlation test. Zoom to Des Moines

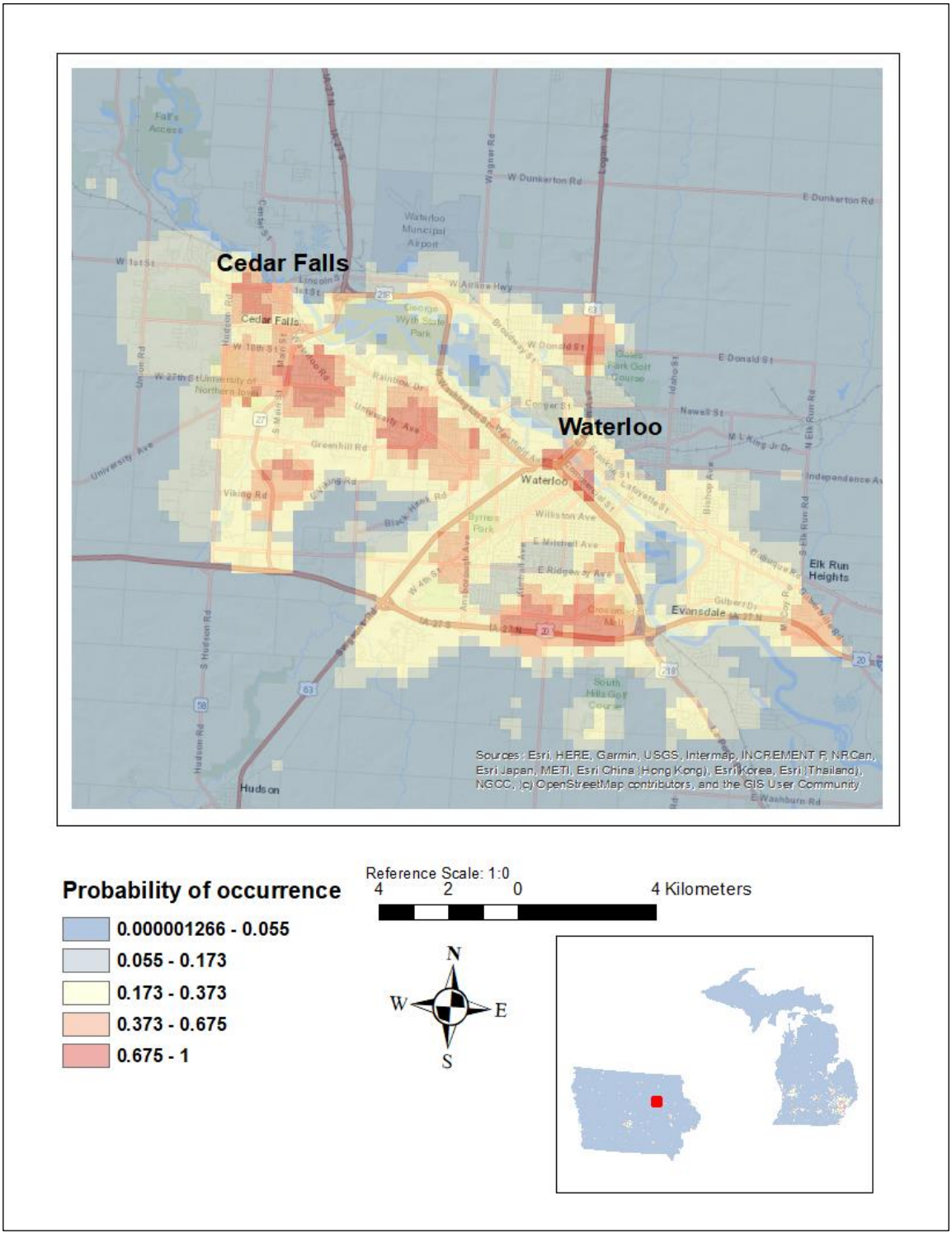


Figure 12. Maxent results after correlation test. Zoom to Cedar Falls

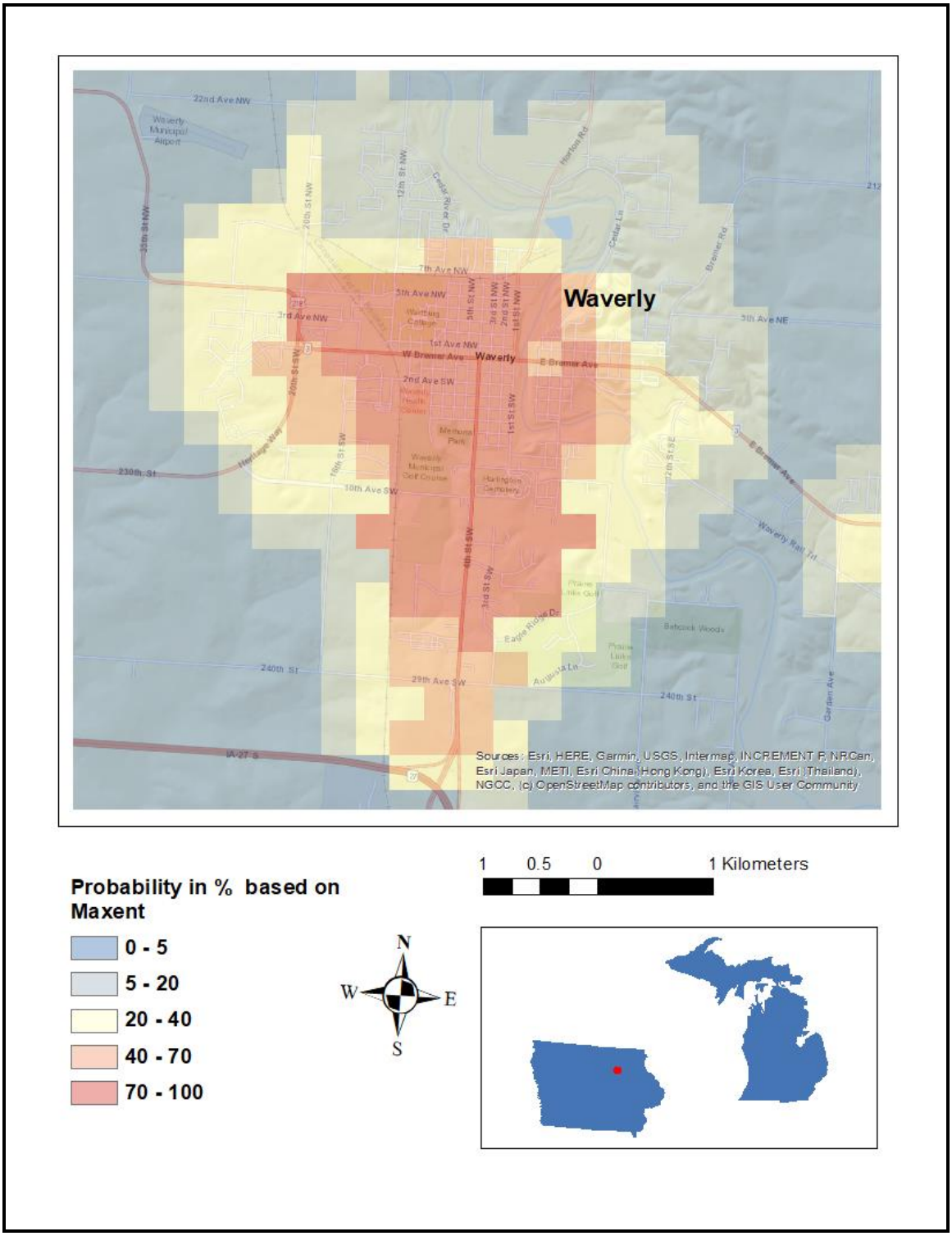


Figure 13. Maxent results after correlation test. Zoom to Waverly

4.5 Model 3. Random Forest Results for Resolution 300 x 300 Meters

After the ecological niche modeling method was used, it was decided to double-check the result using the algorithm of decision trees or random forest. In order to do this, all the location points of Bigbby Cafe were reformatted to a raster, so that a table with a value of 1 (where the cafe exists) and 0 (where the cafe does not exist) was obtained. Further, all the raster cells were reformatted back to dots. And for each current, the raster value of each variable, which coincided with the location of the point, was substituted. Further, using the caret and random forest package in R, a gradation in the significance of each variable for random forest models was revealed.

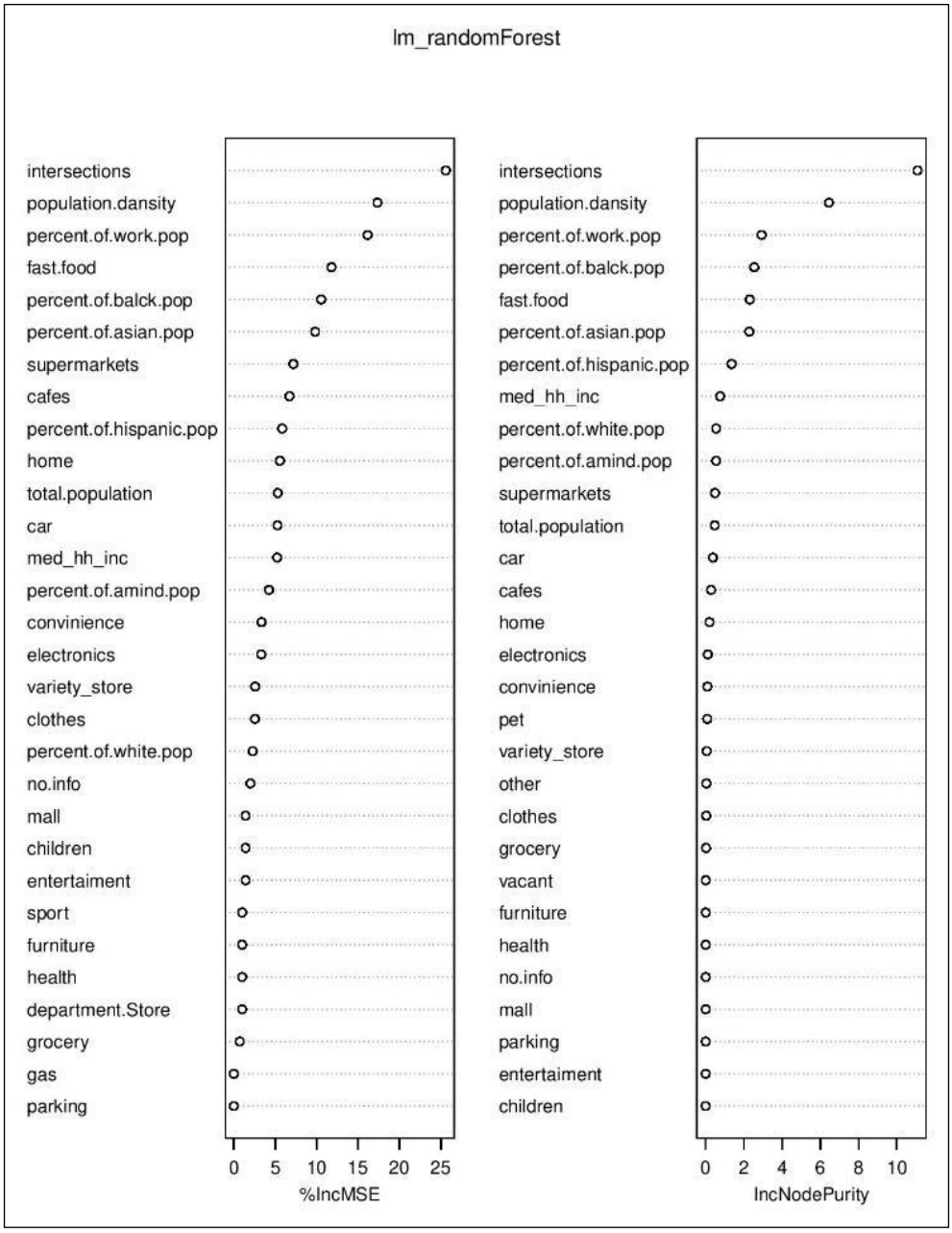


Figure 14. Variable importance for Random Forest model

Unlike Maxent, when applying the random forest algorithm, the most important variables are: intersections density, population density and

percent of work population. Fast food locates only on 4th place, but in the Maxent models it had the largest impact. Five hundred trees were used for calculation. In each split were used 11 variables. Mean of squared residuals was 0.04480722 and this model explained 73.32% of variables. For the model 199 Biggby locations were used as value 1 (exist) and 201 random points inside Michigan state polygon were generated to provide points with value 0 (not exist). The total amount of points was divided for training and testing sets as 80% for training points and 20% for testing points. The accuracy of the model is 98% for training set of points and 96% for testing points.

Accuracy for training set			**Accuracy for testing set**		
	0	1		0	1
0	44	2	0	16	1
1	2	164	1	1	36
Accuracy 98%			Accuracy 96%		

Table 19. Random forest accuracy

4.6 Visual Assessment of Model 3

Visually, you can see that the random forest made a fairly similar classification to Maxent. However, within zones in a higher probability, Maxent made stronger delineations. Random forest, however, designated almost all urban areas as places with very high suitability.

For Des Moines, the forecast by Random Forest is not very accurate. Almost the entire urban area is covered by a zone with a high suitability of the territory. This result is significantly different from what was obtained with the help of logistic regression and Maxent.

We can see in Cedar Falls. Random forest didn't highlight the University Avenue area. However, we see that applying the decision tree methodology works to analyze locations for a business at a macro level.

For Wavery, the forecast is also not very precise. You can see that the entire territory of the city is covered by a zone with a high degree of probability. One for the same resolution, Maxent also made a rather weak prediction.

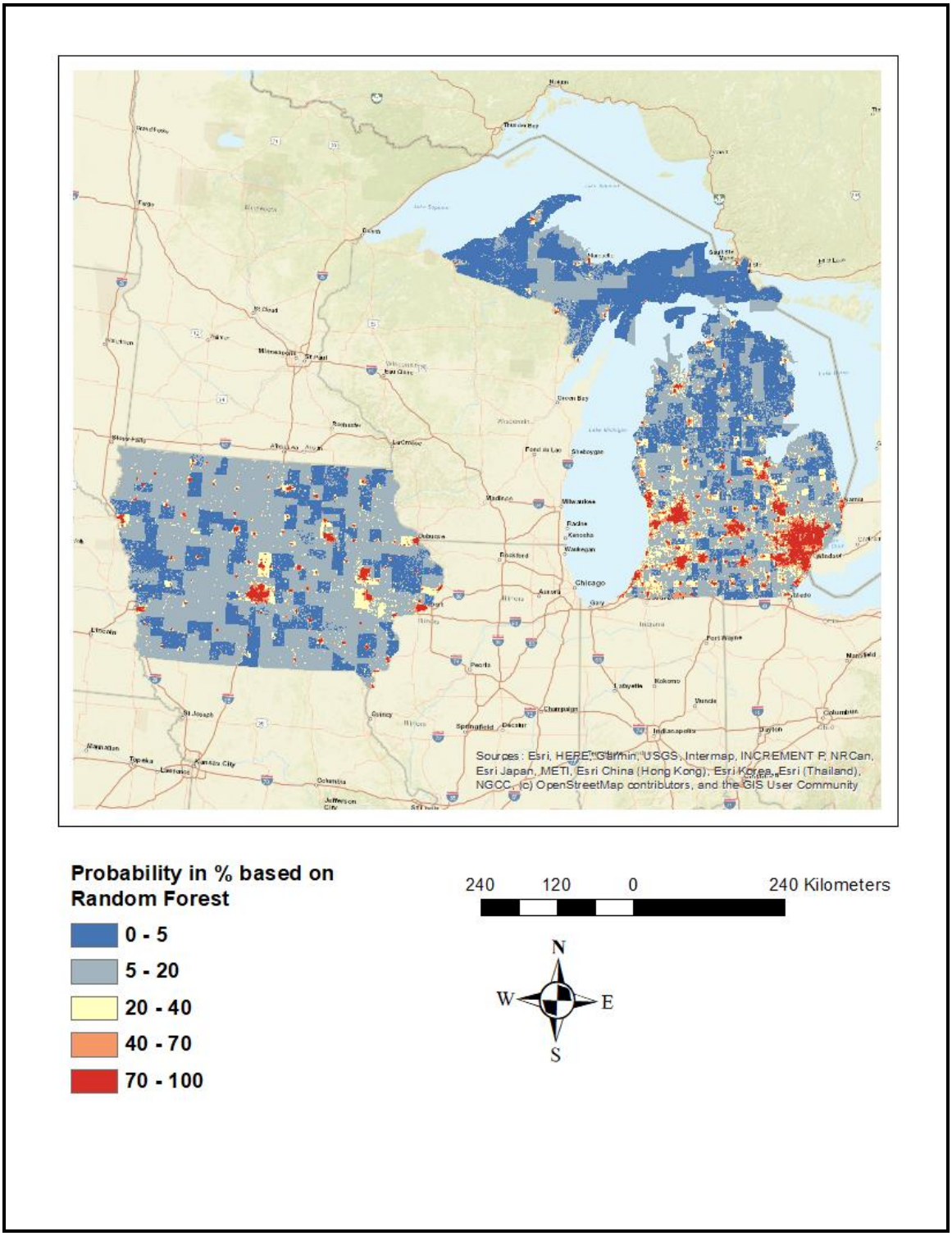


Figure 15. Random Forest classification result for Iowa and Michigan

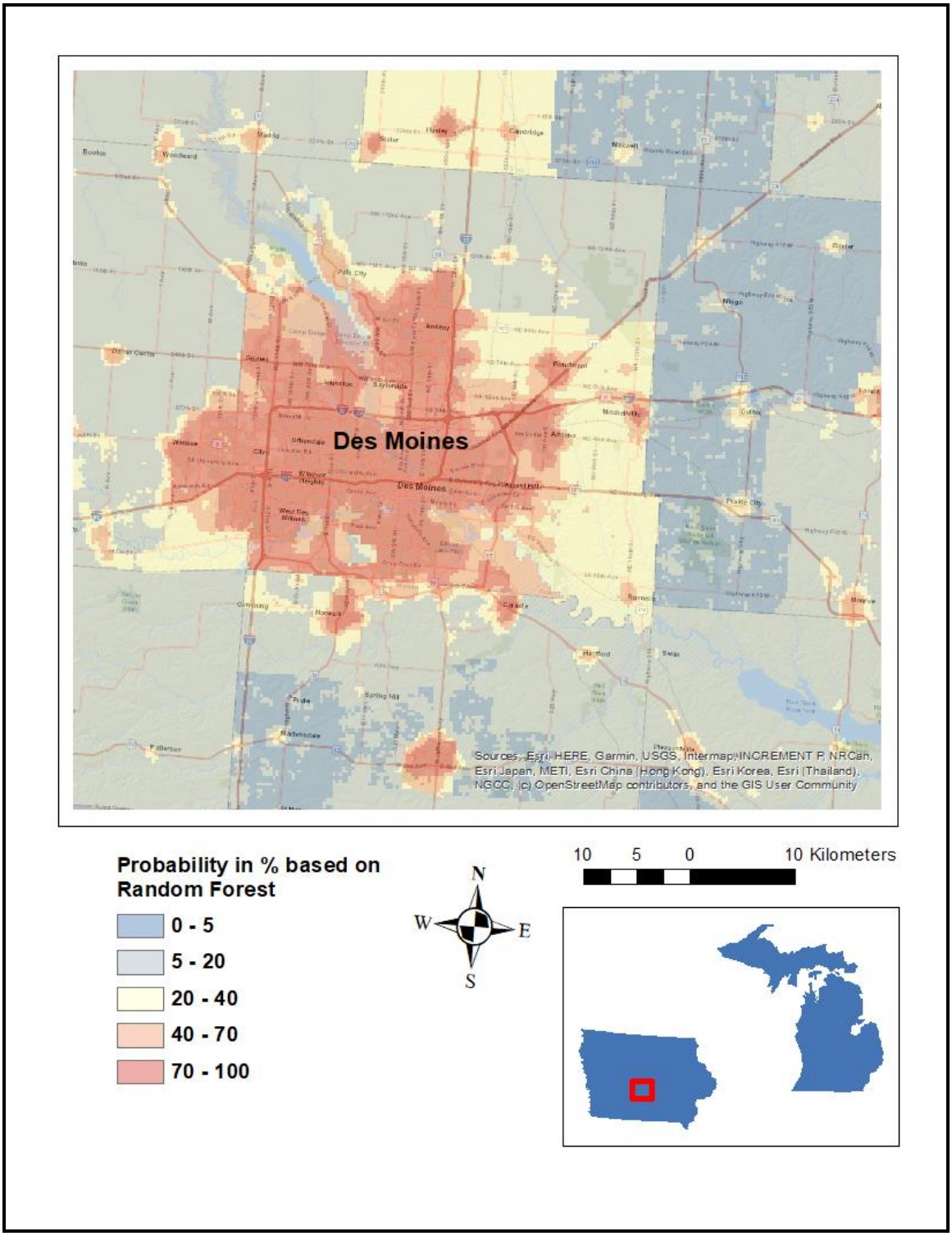


Figure 16. Random Forest classification result for Des Moines

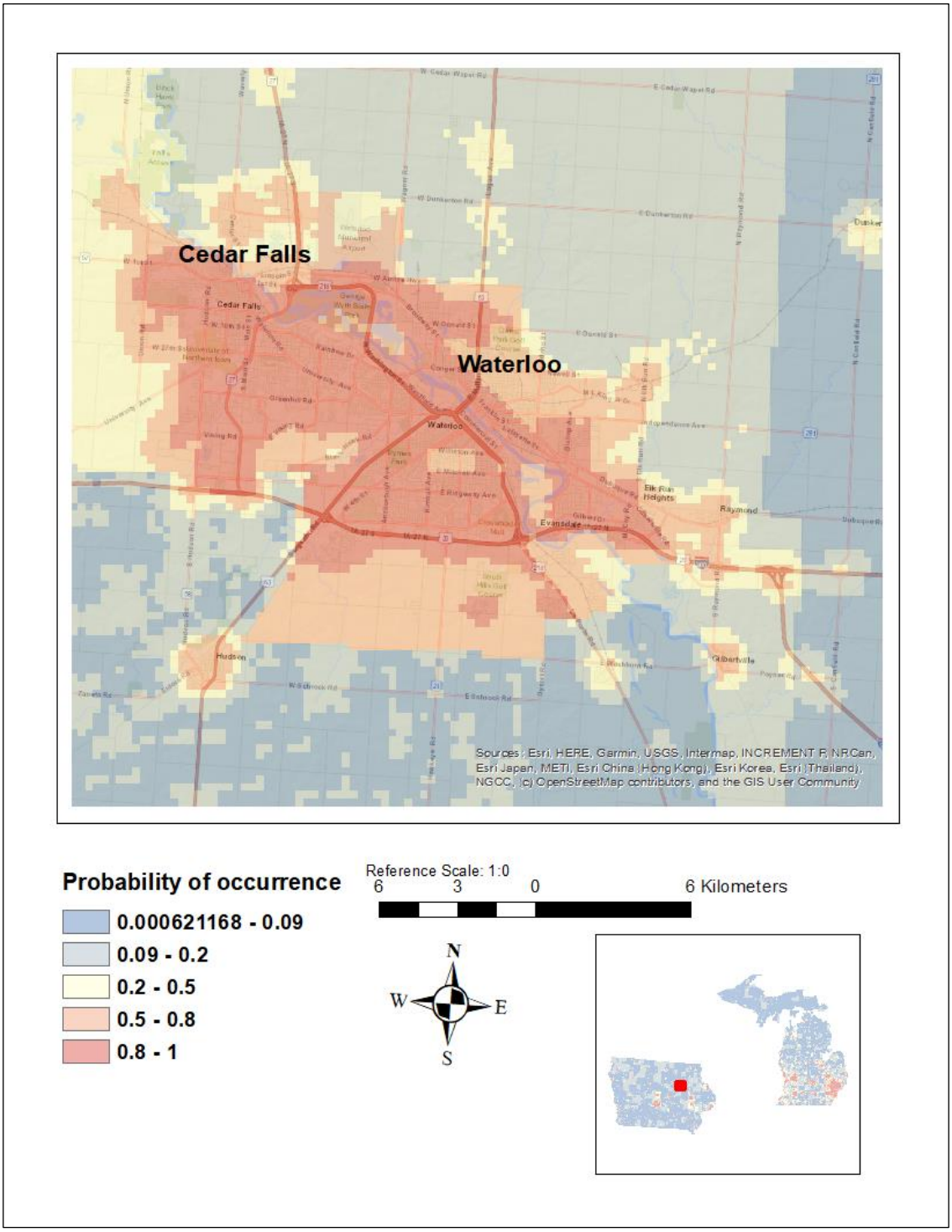


Figure 17. Random Forest classification result Cedar Falls, Iowa

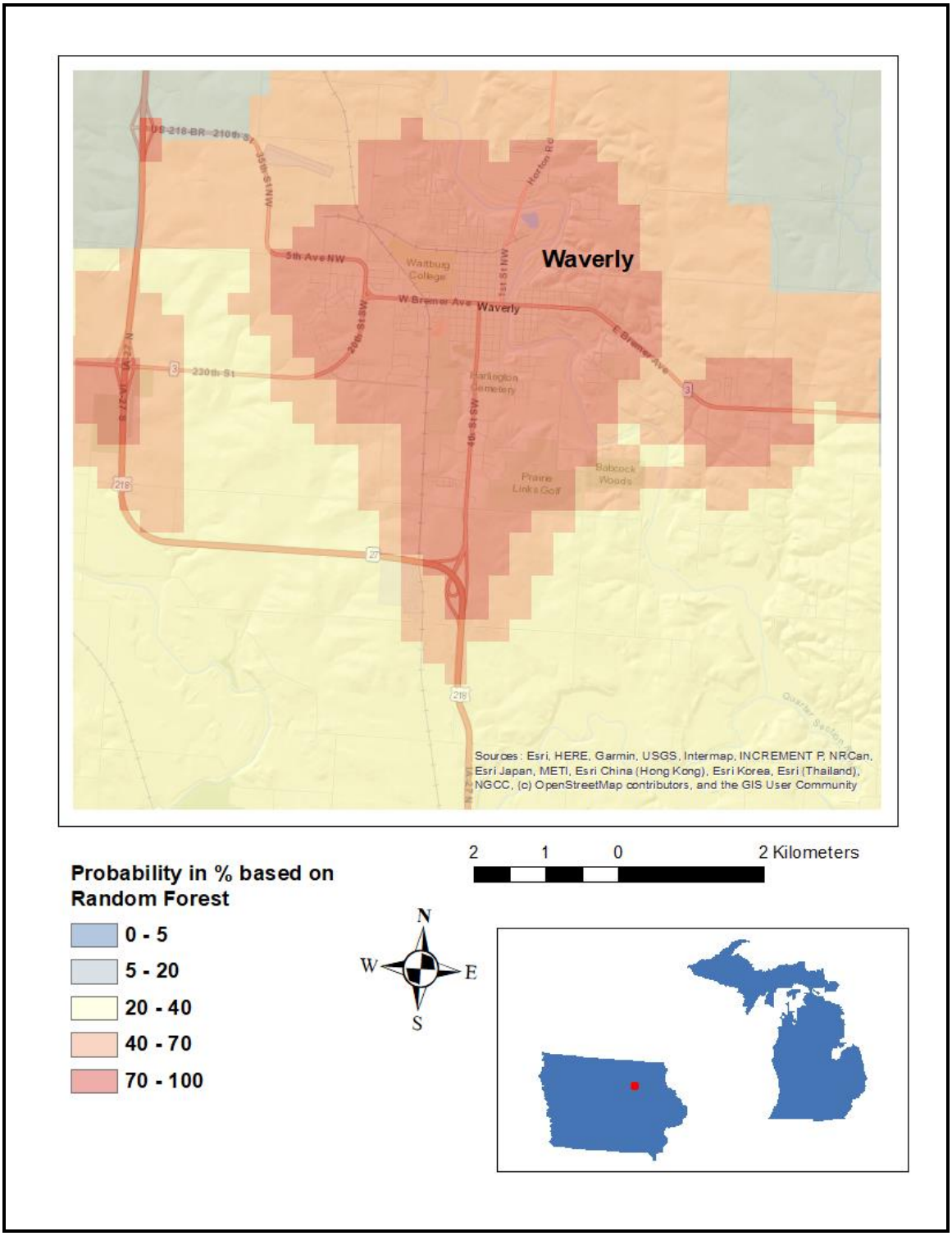


Figure 18. Random Forest classification result Waverly, Iowa

4.7 Difference in Prediction between Maxent and Random Forest

In order to better understand how the results of the forecast of Maxent (Models 1 and 2) and Random Forest (Model 3) differ, the difference between the probability of the forecast of Maxent and the forecast of Forest Forest (Maxen - Rand Forest) was calculated. Thus, we see that at the state level, the predictions are quite identical for rural zones. However, if you look at urban areas, the results differ quite strongly.

So for Des Moines, Random Forest gave a high probability for almost the entire city. For downtown and western Des Moines, Random Forest and Maxent received almost the same probability.

For Cedar Falls, Random Forest also calculated a larger area with a higher probability than Maxent. However, for University Avenue and the East Viking Road shopping area, the predictions match. For Waverly, the same forecast was obtained for almost the entire city, but it is definitely quite small. Random Forest also predicted a high probability outside of Waverly

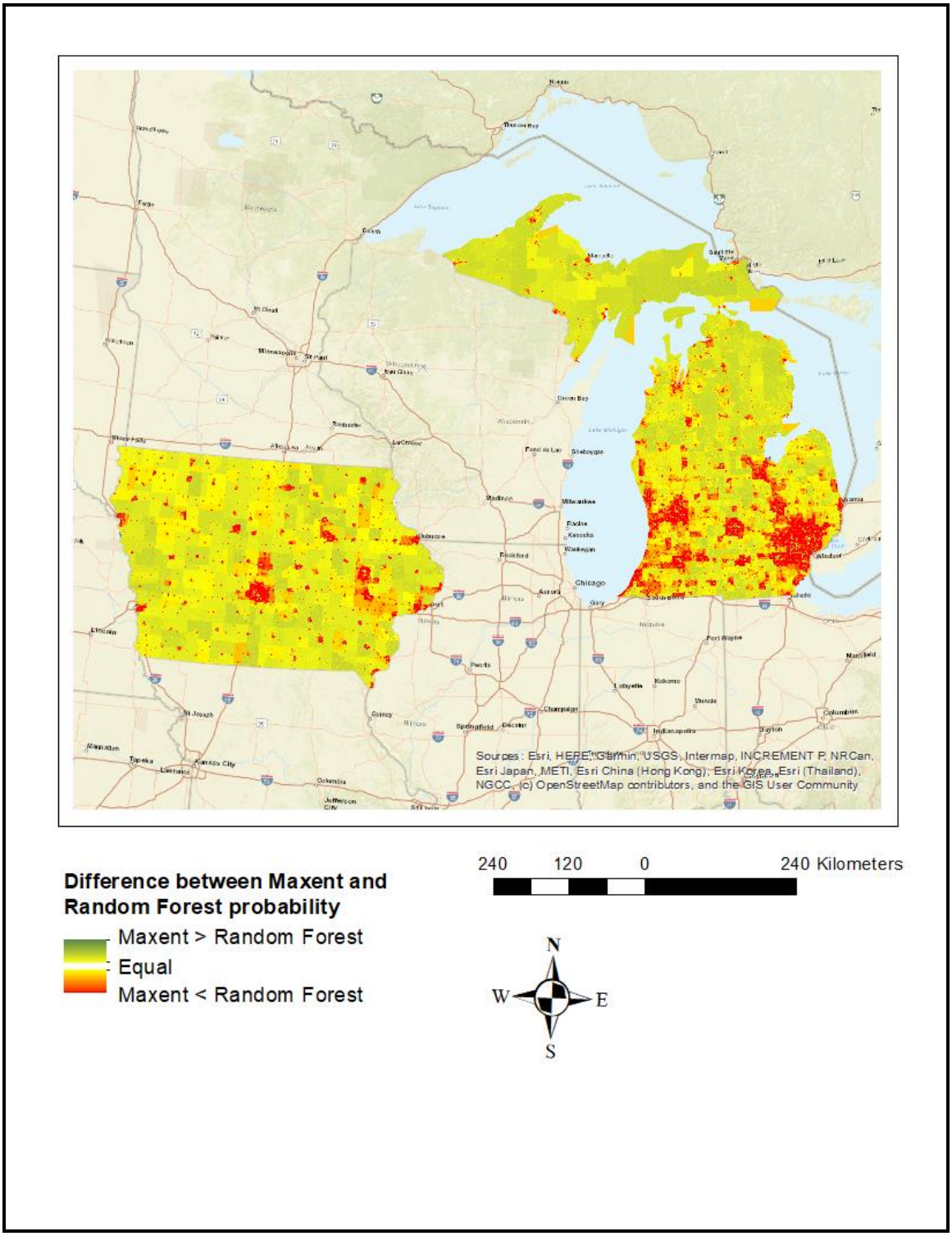


Figure 19. Difference between Maxent and random Forest prediction

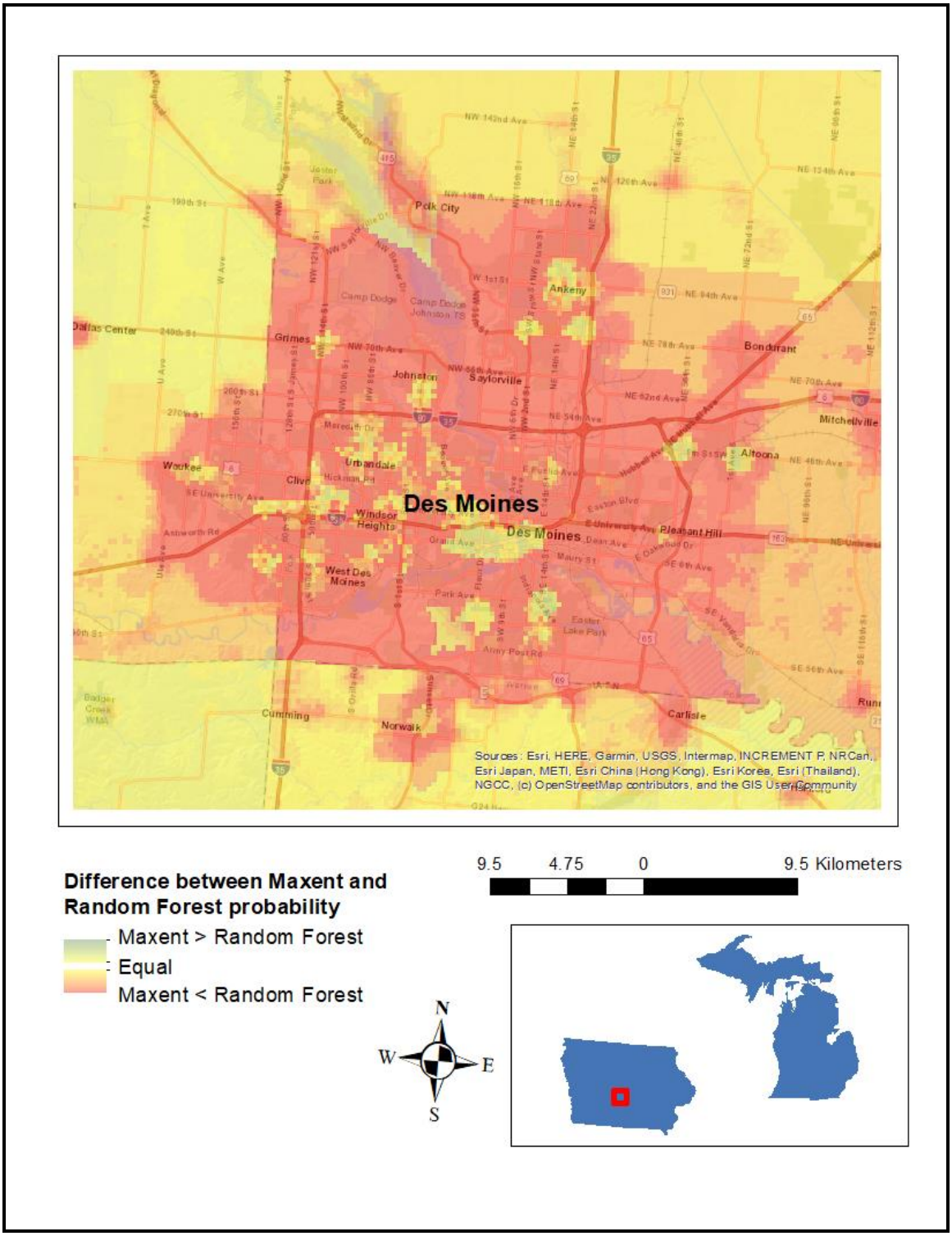


Figure 20. Difference between Maxent and random Forest prediction for Des Moines

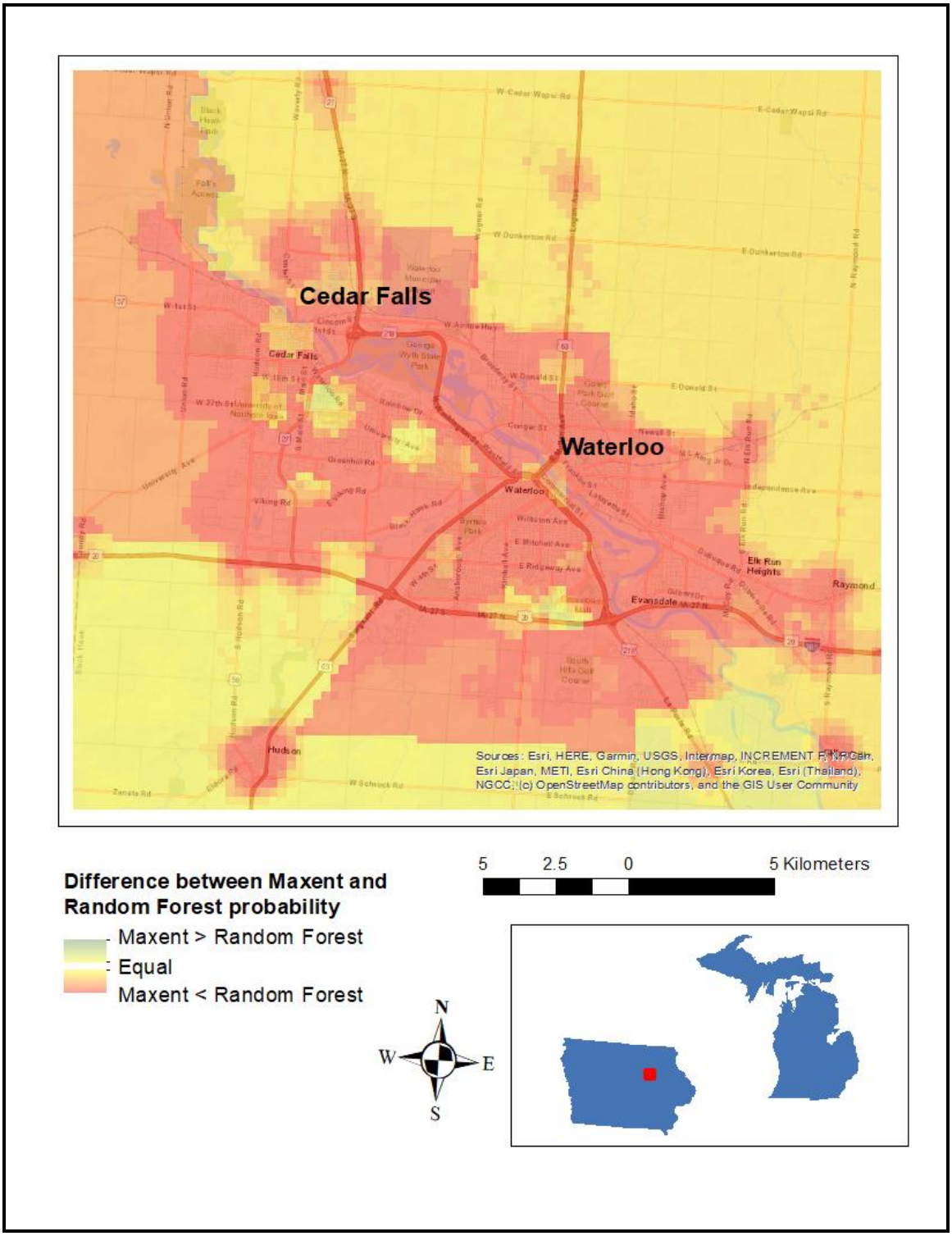


Figure 21. Difference between Maxent and random Forest prediction for Cedar Falls

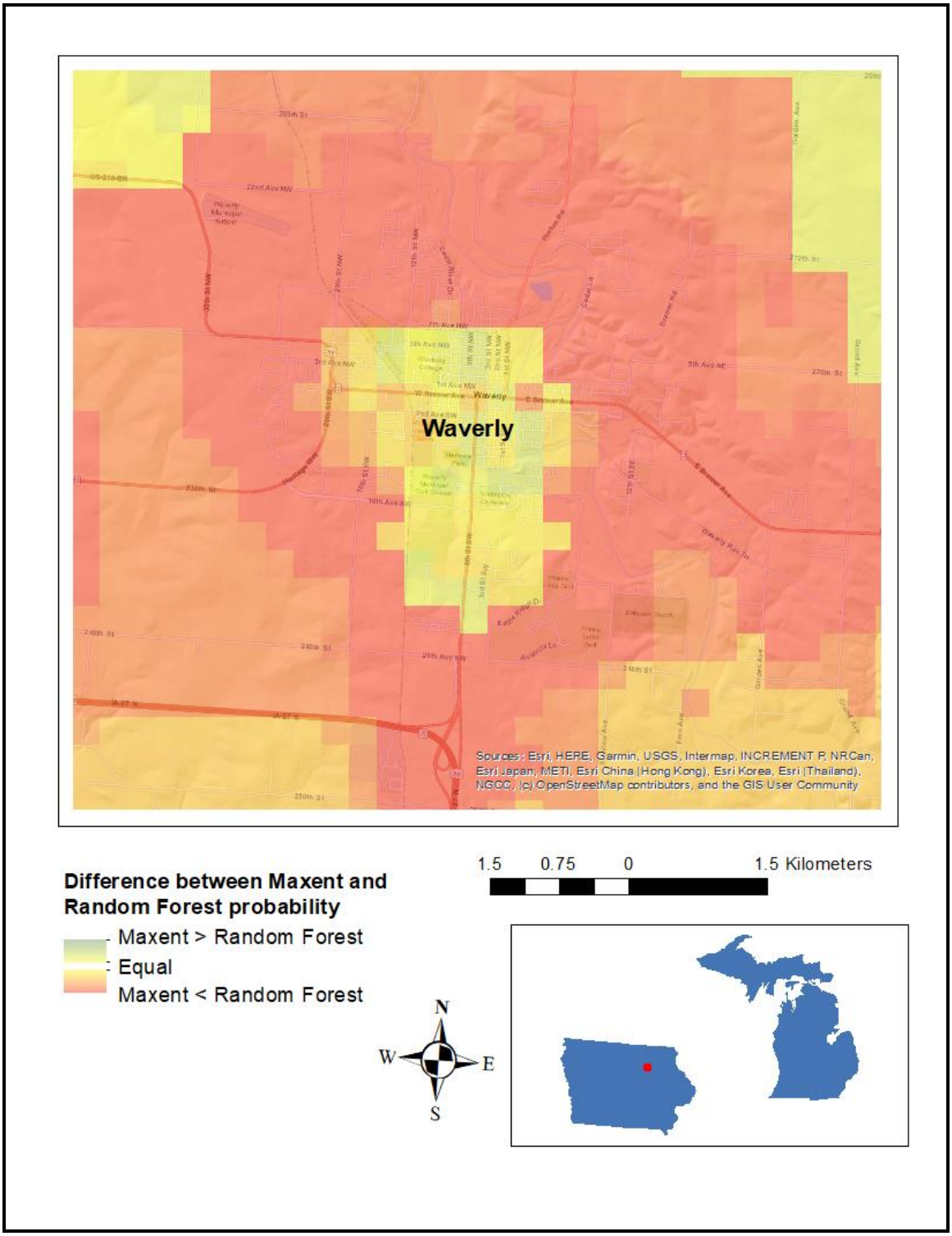


Figure 22. Difference between Maxent and random Forest prediction for Waverly

4.8 Summary

1. Machine learning can be used in business geography and for strategic business planning. So, logistic regression showed a good result. The accuracy of the result was more than 90%, while the number of points for training was only 199, and the total number of lines for prediction was more than 3 million. Random Forest also showed its suitability in this type of problem, however, it is necessary to more accurately assess the required number of points with different classes.

2. For a better classification, a thorough analysis of the variables that will be used in the model is necessary. For both models, more than 30 variables were used, however, only 4-5 of the entire list turned out to be the most significant for them. Thus, it is mono to conclude that it is always necessary to evaluate the quality of a variable and its contribution to the overall result before including it in the model.

CHAPTER 5

DISCUSSION

This study highlighted several new methodological improvements:

1. Creation of a new model for finding locations for business, which is a fairly new approach in the framework of the application of machine learning.
2. Results of covariance checking for variables.
3. What resolution of prediction is better for machine learning algorithm implementation.
4. Description of the results depending on the scale.
5. Impact of research to geomarketing

5.1 Creation of a New Model for Finding Locations for Business, which is a Fairly New Approach in the Framework of the Application of Machine Learning.

There are not so many studies that use machine learning to find locations for a business. However, there are studies on finding places for hotels, restaurants and forecasting the cost of housing. Most of the studies use classic machine learning methods. So, for example, Helber in his article on predicting the cost of housing used Convolutional Neural Network and Random Forest algorithm (Helber et al. 2019). In my research, I decided to use a method that was previously used in other scientific fields. The method of modeling ecological niches was not

previously used to find locations for business. This method is used for scientific purposes to search for ranges of species. However, I did not limit myself to just Maxent, but I also took the standard machine learning algorithm - Random Forest, to understand the difference between the two results. In addition, in other studies, the authors used criteria that were not based on the business model of a particular restaurant or company. For the most part, variables were used that can only give an idea of a search for a location. So, Shihab used data from Yelp. Of course, he took a lot of variables, but perhaps this did not give a pretty picture. In the results, he writes that many algorithms did not give the expected accuracy (Shihab et al. 2018). I in other hand Tugce Bilen took only social economic factors for analysis. In this case, such parameters as the presence of other restaurants or the presence of competitors and population density were also not used (Bilen et al. 2018). In my case, I made a more complex approach. I proceeded from Biggby business models. Such an approach is, in my opinion, the closest to reality and allows in the future to use it for strategic planning. Of course, it is impossible to find all the necessary data for educational purposes. I used not only the socio-demographic data that characterize the population, but also an indicator of the density of competitors and other commercial organizations that can facilitate the opening of coffee houses. In addition, I used the data on the number of

crossings per unit of measurement. This approach broader characterizes the process of finding locations.

5.2 Results of Covariance Checking for Variables

One of the interesting results was the analysis of covariance. For my model, I selected 38 variables and in order to check how much they correlate with each other. After the covariance test, it turned out that only 3 variables from the entire list showed that they have a correlation with other variables. However, even when searching for variables that correlated among themselves, a difference arose. So, when calculating the coefficient of Spearmen, it was found that the variable number of restaurants per unit area correlated with the variable number of services (index = 0.641). Restaurants also correlated with fast food (index = 0.706). The service variable correlated with a variable with respect to the number of clothing stores and with a variable that shows the number of car stores. However, according to the results of analysis of the covariance of the VIF, only variables of Median Household income and Average Disposable income were detected as autocorrected.

James Gareth in his book “An Introduction to Statistical Learning” says that in statistics, the variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone (James et al. 2013). This way we check how regression changes if we change one of the variables. Thus, for our variables, 2

variables had a high degree. Gregory Colder in his book he describes that the Spearman coefficient characterizes how the value of a variable changes, when changing the value of another variable. If Y tends to increase when X increases, the Spearman correlation coefficient is positive. If Y tends to decrease when X increases, the Spearman correlation coefficient is negative (Corder and Foreman 2014). It was a surprise for me that most of the variables that characterize the amount of a business per unit of measure do not correlate with each other. However, the articles by Helber (Helber et al. 2019) and Bilen (Bilen et al. 2018), which focus on the use of machine learning to find locations for business, do not contain information about checking data for collinearity. Perhaps this is due to the fact that they did not take so many variables and there was no need to check for a correlation between them. However, Young used the projection pursuit regression (PPR) model for her research (Yang et al. 2015).

5.3 What Resolution of Prediction is better for Machine Learning Algorithm Implementation.

Another point that is debatable is the resolution or pixel size, on the basis of which the calculation is appropriate. For my work, first of all, I fixed the size of the pixel equal to 100 per 100 meters. This size is one section in a block. So I wanted to do an analysis that could show the result at the level of each house. Then I took a pixel size of 300 by 300 meters, which is the size of a single bar. Young in her work took a pixel size equal

to 200 by 200 meters, but her analysis took place within the framework of the city of Beijing. Her article did not indicate why it was she who decided to take a pixel size of that size (Yang et al. 2015). Helber in his work used images with a resolution of 10 cm per pixel and used them at different scales. His work was applied on the scale of Amsterdam, so this permission is fully justified. Then all the pictures were converted to a size of 256 by 256 pixels, which meant that about 40 cm was per pixel (Helber et al. 2019). Thus, the question arises of what pixel resolution should be taken for analysis using machine learning. I think that the most accurate approach here would be to assess the scale of the zone that needs to be explored. It is also necessary to understand what specific goal the company will pursue when conducting such a study. If the study is necessary to conduct strategic planning and assess which city should open the next point, then the pixel size can be reduced to kilometers. If we are talking about researching a specific city and choosing a district or a specific street, then the pixel size should be within hundreds of meters.

5.4 Description of the Results Depending on the Scale.

In this paper, two methods were used to calculate the suitability of locations for opening a specific business. Maxent uses a logistic regression model as a computational algorithm. For exploring the dataset, a statistical method is used which is called logistic regression. So in another word, logistic regression is the statistical method of evaluating the

particular dataset. Logistic Regression has similarity with the concept of probability. If an event occurs, the probability will be:- P(event)=occurred event/total number of event But in Logistic Regression, the probability will be the event is occurring versus the event is not occurring.

Another method was the decision tree is an algorithm which is used to reach in a decision or to get the target value. Ross Quinlan invented this algorithm. The target values can be set in a tree model called classification tree. In this tree, there are leaves and parents like other trees. Here the leaf nodes represent the class labels and branch represents features of that label. Decision trees where target values can take continuous values is called a regression tree. We use a decision tree to describe the whole decision-making process in a particular way which makes the entire thing easily understandable. Decision tree normally used operation research especially decision analysis (Shihab et al. 2018). Thus, different methods show us different results. Using Cedar Falls as an example, you can clearly see that the logistic regression model did better with prediction than random forest. A change in the number of training points with indices 0 did not lead to an improvement in the result, as well as a change in the number of the number of trees. However, for my work, I used the results of both methods and excluded variables that were identified in one and the other case.

5.5 Impact of Research to Geomarketing

As I noted before in my research. Spatial data now becomes more commonly used in business. Companies already have huge amount of date with spatial binding. This information also can be used not only for developing chain or for finding new locations, but also for understanding notational competitors strategy. Based on existing competitor locations, you can predict their next steps If you understand what kind of business model your competitor has and in what locations the points are located, then you can understand in advance where most likely he will open his next restaurant or store.

With the introduction of 5G of the Internet, companies will be able to receive more accurate data on the locations of their users. More and more devices will have an Internet connection and transmit their data with more accurate data with location parameters. Machine learning can significantly improve the processing of this data, as well as provide an opportunity to better personalize communication with the customer.

Historical customers data also can be very useful to see new patterns of behavior. If it would be possible to combine these kind of data with traffic, history of purchase in grocery stores, home addresses and other variables it would push geomarketing in a new level. The location of infrastructure and commercial properties will become more convenient for the user, as companies will more accurately know their client's behavior

CHAPTER 6

CONCLUSIONS

Today, more and more spheres of human activity are beginning to apply machine learning methods. This happens quite organically, since data is accumulated that can be useful for making new decisions. Machine learning is widely used in medicine, banking, and insurance. The purpose of this work was to show whether machine learning can be used to solve business problems in geography. This study was aimed not only at taking, which could play a role in determining the location, namely those variables that Biggby guides to search for locations for new points. Existing models were built on the basis of data that could be found in the public domain, so the models could not describe the full picture. However, the result of the analysis was a fairly large percentage of accuracy.

6.1 Accuracy

One of the main conclusions is the conclusion that the accuracy of the analysis in both algorithms gives an accuracy of more than 90%. However, a visual analysis shows that the results differ quite significantly. The method of biological modeling of niches using logistic regression yields a result closer to real life and makes more precise allocation of zones. Unlike logistic regression, the random forest model also correctly identifies areas where cafes could most likely be opened, but the accuracy of this algorithm is limited by the scale of cities, in the case of Cedar Falls, or areas in the

case of Detroit or Des Moines. Thus, we can say that in this case, logistic regression works better and makes it possible to determine locations with a fairly high accuracy, even at the macro level. Of course, everything depends on the size of the unit of area, on the basis of which the analysis is built, but in our case, the unit of area fit even in such small urban scales as Cedar Falls. Thus, the logistic regression and the Maxent program have proven themselves as a tool and tool for choosing a location. Radom Forest, in turn, is more suitable for strategic planning, which will allow the company to evaluate in which direction it is worth developing its network or which area should be developed next. Most likely, the Random Forest algorithm will be useful for the process of comparing two cities, as part of the assessment of specific areas. For this case, perhaps there were too few points at which a Random Forest could train its algorithm in order to produce more accurate results.

6.2 Main variables

For this study, 39 variables were selected, based on which the model was built. These variables were selected first based on the Biggby business model. Thanks to this, I did not proceed from general ideas about where the cafe should be, but was guided directly by those parameters that were laid down directly to search for locations of this particular network. However, not all the necessary data were found for this study. So, the traffic intensity could not be found for the entire state, so this data was

not included in the model. However, from the entire list of included variables, 5 variables turned out to be most useful, such as: the number of crossings within a radius of one mile from a pixel, the number of fast food, population density, percentage of working population, and median household income. Thus, only 20% of the variables were useful in designing models. We can say that Paretto's law is observed here, which says that 80% of the result is obtained when 20% of the energy is consumed. In my case, I tried to add to add those variables that were necessary for the business model, but at the same time create as many categories within these variables as possible. Thus, I chose not only catering places, but divided them into cafes, restaurants and fast food. I did the same with shops. I did not take the store category, but divided it and many types. This was done in order to find some new patterns in the results that the Biggby business model does not describe. Thus, we saw that fast food is of the greatest importance among public catering places, and supermarkets among stores.

6.3 Resolution

If we talk about resolution, then according to the results of our work, we can say that the use of machine learning makes it possible to do analysis for strategic planning. The result of this work showed that even applying the algorithm to such a large territory as several states in the USA, it is possible to obtain a satisfactorily accurate result, at least by the location

of the zones in which you need to check for the possibility of placing a cafe there. For a more accurate result, it is already necessary to specifically apply the algorithm based on data of a smaller scale. For example, it is necessary to compare cities of similar size with more accurate data and apply the algorithm directly to them. Also, due to scale, the number of points for training should be larger in the case of Random Forest.

6.4 Limitations

This work met with several limitations. First of all, this lack of information directly from Biggby company itself. As part of my company, I wanted to get in touch with companies in order to learn more about their methods of choosing new places for their cafes. Unfortunately, the company employees did not give any answers to my questions, so I had to find information about their business model on the Internet. Thus, the information about the Biggby business module, and accordingly the information about the main variables, could be inaccurate.

The second limitation, as in many similar works, is the limitation on the availability of open data. For my work, I used ESRI data on socio-demographic indicators, as well as on population incomes. I also used the Open Street Map data to locate all the stores, catering places, and crossings. However, there is open intensity data. Traffic or just general data on the road network with data on high-speed mode could not be found. In addition, all the data that was collected needed to be structured.

Therefore, most of the work was not done by research or hypothesis construction, but by preparing the data and bringing them to the right form. For example, it was very difficult to find socioeconomic data for all 4 states. As a result, I bought them from ESRI. If a database appears that has socio-economic data, traffic data and current company data, then machine learning could get a very blue push. However, most companies will not share their data. This can be understood. In this regard, most of the works related to data analysis for commercial organizations are limited by the lack of actual data.

Another limitation has become too much data. Due to the fact that for my work it was necessary to have data for 4 states with a pixel size of 100 by 100 meters, each variable was in a table with more than 50 million rows. And because 39 variables were needed for my work, the weight of the data milestones was more than 2 GB. Not a single computer could calculate such a volume of data, or the calculation took more than 5 hours. In connection with these, it was decided to reduce the number of states and increase the number of pixels to 300 by 300 meters. This led to the possibility of calculating the data and using them in R studio. Thus, for a more accurate result, a more powerful computer is needed, as well as the creation of a database that could prevent overloading RAM.

6.5 Future directions

It would be most logical to try using other machine learning techniques to find locations for business. In the future, I would also like to work more closely with companies and network development managers to obtain more accurate results. Such cooperation could lead to a more efficient use of machine learning to find locations for business, due to the availability of more relevant data directly from the company. The next step in my work will be not only an analysis of the methodology for predicting locations, but also a method for checking how accurate the predictions are and how the company can evaluate the potential of each point in monetary terms.

REFERENCES

"25 Best Coffee Franchises of 2020 (Updated Rankings)." 2017. *Franchise Chatter* (blog). November 24, 2017. https://www.franchisechatter.com/2017/11/23/the-25-best-coffee-franchises-of-20172018/.

"Annual Report John Deere." 2017. United States Securities and Exchange Commission.

Aversa, Joseph M. 2019. "Spatial Big Data Analytics: The New Boundaries of Retail Location Decision-Making." PhD diss., Wilfrid Laurier University. https://scholars.wlu.ca/etd/2138

Baker, Robert. 2002. "Regions of Time and Internet: Modelling: An Application of the Space-Time Trip (RASTT) Model to the USA Internet Market." Paper presented at the 42nd Congress of the European Regional Science Association: "From Industry to Advanced Services Perspectives of European Metropolitan Regions," Dortmund, Germany, Louvain-la-Neuve, August 27th - 31st, 2002.

Benito, B Pando de, and J Giles de Peñas. 2008. "Greenhouses, Land Use Change, and Predictive Models: MaxEnt and Geomod Working Together." In *Modelling Environmental Dynamics*, edited by Martin Paegelow and María Teresa Camacho Olmedo, 297–317. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68498-5_11.

Bensic, Mirta, Natasa Sarlija, and Marijana Zekic-Susac. 2005. "Modelling Small-Business Credit Scoring by Using Logistic Regression, Neural Networks and Decision Trees." *Intelligent Systems in Accounting, Finance and Management* 13 (3): 133–50. https://doi.org/10.1002/isaf.261.

"BIGGBY® COFFEE Locations." 2020. BIGGBY COFFEE®. January 20, 2020. https://www.biggby.com/locations/.

"BIGGBY COFFEE Cuts Franchise Start-Up Costs 40%." 2011. BIGGBY COFFEE®. July 1, 2011. https://www.biggby.com/biggby-coffee-cuts-franchise-start-up-costs-40/.

"BIGGBY COFFEE Franchise Review on Top Franchise Opportunity Blog." 2011. *Franchise Chatter* (blog). August 8, 2011.

https://www.franchisechatter.com/2011/08/08/biggby-coffee-develops-new-store-footprint-reducing-start-up-costs-by-nearly-40-to-210000/.

Bilen, Tugce, Muge Erel-Ozcevik, Yusuf Yaslan, and Sema F. Oktug. 2018. “A Smart City Application: Business Location Estimator Using Machine Learning Techniques.” In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 1314–21. Exeter, United Kingdom: IEEE. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00219.

Birkin, Mark, Graham Clarke, and Martin P. Clarke. 2002. *Retail Geography and Intelligent Network Planning*. John Wiley & Sons.

Breiman, Leo. 1999. “Random Forests.” 1999. http://machinelearning202.pbworks.com/w/file/fetch/60606349/breiman_randomforests.pdf.

———. 2001. “Random Forests.”

Byrom, J.W. 2005. “The Use of Data in Outlet Locational Planning: A Preliminary Ex Amination across Retail and Service Sectors.” *Management Research News* 28 (5): 63–74. https://doi.org/10.1108/01409170510629014.

Chen, Ji, Jinsheng Wang, Tomas Baležentis, Fausta Zagurskaitė, Dalia Streimikiene, and Daiva Makutėnienė. 2018. “Multicriteria Approach towards the Sustainable Selection of a Teahouse Location with Sensitivity Analysis.” *Sustainability* 10 (8): 2926. https://doi.org/10.3390/su10082926.

Clarke, Graham. 1998. “Changing Methods of Location Planning for Retail Companies.” *GeoJournal* 45: 289–298.

“Collinearity and Stepwise VIF Selection – R Is My Friend.” 2013. https://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/.

Corder, Gregory W., and Dale I. Foreman. 2014. *Nonparametric Statistics: A Step-by-Step Approach*. John Wiley & Sons.

Eric Decker. 2016. “Blog 2.2 – Biggby Coffee.” *Eric Decker* (blog). January 22, 2016. https://ericdeckerblog.wordpress.com/2016/01/22/blog-2-2/.

France, Stephen L., and Sanjoy Ghose. 2019. “Marketing Analytics: Methods, Practice, Implementation, and Links to Other Fields.” *Expert Systems with Applications* 119 (April): 456–75. https://doi.org/10.1016/j.eswa.2018.11.002.

Fujisawa, Shizuka. 2013. “Pedestrian Counting in Video Sequences Based on Optical Flow Clustering.” *International Journal of Image Processing (IJIP)* 7 (1): 16.

Greenacre, Michael, and Raul Primicerio. 2008. “Measures of Distance between Samples: Euclidean.” Fundacion BBVA Publication (December 2013). http://www.econ.upf.edu/~michael/stanford/maeb4.pdf.

Helber, Patrick, Benjamin Bischke, Qiushi Guo, Jorn Hees, and Andreas Dengel. 2019. “Multi-Scale Machine Learning for the Classification of Building Property Values.” In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 4873–76. Yokohama, Japan: IEEE. https://doi.org/10.1109/IGARSS.2019.8900257.

Hernández, Tony, and David Bennison. 2000. “The Art and Science of Retail Location Decisions.” *International Journal of Retail & Distribution Management* 28 (8): 357–67. https://doi.org/10.1108/09590550010337391.

Hernandez, Tony, and Marco Biasiotto. 2001. “Retail Location Decision-Making and Store Portfolio Management.” *Canadian Journal of Regional Science* 24 (3): 399.

Hess, Ronald L., Ronald S. Rubin, and Lawrence A. West. 2004. “Geographic Information Systems as a Marketing Information System Technology.” *Decision Support Systems* 38 (2): 197–212. https://doi.org/10.1016/S0167-9236(03)00102-7.

Hutchinson, Evelyn. 1957. “Concluding Remarks.” Yale University, New Haven, Conecticut.

Iowa Workforce Development. 2004. “Major Industries in Iowa.” In *Iowa: Life Changing*.

Iowa Workforce Development, Finance & Insurance Iowa Industry Profile. 2018. “Finance & Insurance Iowa Industry Profile.” IOWA Workforce Development. https://www.iowaworkforcedevelopment.gov/labor-market-information-division.

Iowa Workforce Development, Health care and social assistance Iowa Industry Profile. 2018. “Health Care and Social Assistance, Iowa Industry Profile.” 2018. https://www.iowaworkforcedevelopment.gov/sites/search.iowaworkforcedevelopment.gov/files/documents/2018/healthcare_2018_0.pdf.

Iowa Workforce Development, Real estate and leasing Iowa Industry Profile. 2018. “Real Estate and Rental and Leasing, Iowa Industry Profile.” 2018. https://www.iowaworkforcedevelopment.gov/sites/search.iowaworkforcedevelopment.gov/files/documents/2018/realestate_2018_0.pdf.

Irizarry, Javier, Ebrahim P. Karan, and Farzad Jalaei. 2013. “Integrating BIM and GIS to Improve the Visual Monitoring of Construction Supply Chain Management.” *Automation in Construction* 31 (May): 241–54. https://doi.org/10.1016/j.autcon.2012.12.005.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.

Jaynes, E. T. 1957. “Information Theory and Statistical Mechanics.” *Physical Review* 106 (4): 620–30. https://doi.org/10.1103/PhysRev.106.620.

Johnston, Kevin M, and Elizabeth Graham. 2013. “Spatial Analyst - Suitability Modeling.” Technical Workshop presented at the ESRI International User Conference, San Diego, California, July.

Jones, Ken, and Tony Hernandez. 2004. “Retail Applications of Spatial Modelling.” In *Applied GIS and Spatial Analysis*, edited by Graham Clarke, 11–33. Chichester, West Sussex, England ; Hoboken, NJ: Wiley.

Lao, Yong. 1993. “Solving Large-Scale Location Spatial Interaction Models for Retail Analysis: A GIS Supported Heuristic Approach.” Phd diss., The Ohio State University.

Lee, S. 2005. “Application of Logistic Regression Model and Its Validation for Landslide Susceptibility Mapping Using GIS and Remote Sensing Data.” *International Journal of Remote Sensing* 26 (7): 1477–91. https://doi.org/10.1080/01431160412331331012.

Lin, Xiangyi, and Yuanyuan Zu. 2013. “Multi-Criteria GIS-Based Procedure for Coffee Shop Location Decision,” 47.

Linder G. Ringo. 2009. “Utilizing GIS-Based Site Selection Analysis for Potential Customer Segmentation and Location Suitability Modeling to Determine a Suitable Location to Establish a Dunn Bros Coffee Franchise in the Twin Cities Metro, Minnesota.” http://www.gis.smumn.edu/ GradProjects/RingoL.pdf.

Liu, Tianshun. 2012. “Combining GIS and the Huff Model to Analyze Suitable Locations for a New Asian Supermarket in the Minneapolis and St. Paul, Minnesota USA.” Department of Resource Analysis, Saint Mary’s University of Minnesota, Winona, MN 55987.

Long, J. Scott. 1997. “Regression Models for Categorical and Limited Dependent Variables" (Vol. 7). In *Advanced Quantitative Techniques in the Social Sciences*. Thousand Oaks, CA: Sage.

Longley, Paul A., and Graham Clarke. 1995. *GIS for Buisness and Service Planning*. John Wiley & Sons. http://elibrary.dephub.go.id/elibrary/media/catalog/0010-081500000000007/swf/413/GIS%20for%20Buisness%20and%20service%20planning0001.PDF.

Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill Series in Computer Science. New York: McGraw-Hill.

Okabe, Atsuyuki, Ken Aoki, and Wataru Hamamoto. 1986. “Distance and Direction Judgment in a Large-Scale Natural Environment: Effects of a Slope and Winding Trail.” *Environment and Behavior* 18, no. 6 (1986): 755-772.

Okabe, Atsuyuki, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. 2009. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons.

Okabe, Atsuyuki, and Kei-ichi Okunuki. 2001. "A Computational Method for Estimating the Demand of Retail Stores on a Street Network and Its Implementation in GIS." *Transactions in GIS* 5 (3): 209–20. https://doi.org/10.1111/1467-9671.00078.

Parisien, Marc-André, and Max A. Moritz. 2009. "Environmental Controls on the Distribution of Wildfire at Multiple Spatial Scales." *Ecological Monographs* 79 (1): 127–54. https://doi.org/10.1890/07-1289.1.

Park, Kunsoon, and Mahmood A. Khan. 2006. "An Exploratory Study to Identify the Site Selection Factors for U.S. Franchise Restaurants." *Journal of Foodservice Business Research* 8 (1): 97–114. https://doi.org/10.1300/J369v08n01_07.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology* 49 (12): 1373–79. https://doi.org/10.1016/S0895-4356(96)00236-3.

Petrov, Andrey N., and Jordan M. Wessling. 2015. "Utilization of Machine-Learning Algorithms for Wind Turbine Site Suitability Modeling in Iowa, USA: Machine-Learning Algorithms for Wind Turbine Site Suitability." *Wind Energy* 18 (4): 713–27. https://doi.org/10.1002/we.1723.

Phillips, Steven J., Robert P. Anderson, Miroslav Dudík, Robert E. Schapire, and Mary E. Blair. 2017. "Opening the Black Box: An Open-Source Release of Maxent." *Ecography* 40 (7): 887–93. https://doi.org/10.1111/ecog.03049.

Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190 (3–4): 231–59. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Phillips, Steven J, and Miroslav Dudık. 2007. "Modeling of Species Distributions with Maxent: New Extensions and a Comprehensive

Evaluation." *Ecography 31: 161-175, 2008*, 15. https://doi.org/10.1111/j.2007.0906-7590.05203.x.

Phillips, Steven J, and T Research. 2017. "A Brief Tutorial on Maxent," 39.

Pick, James B. 2008. *Geo-Business: GIS in the Digital Organization*. John Wiley & Sons.

Polyakov, L. 1971. "Spearman's Rank Correlation Coefficient." *M.: ЮНИТИ*, 56.

Ponder, W. F., G. A. Carter, P. Flemons, and R. R. Chapman. 2001. "Evaluation of Museum Collection Data for Use in Biodiversity Assessment." *Conservation Biology* 15 (3): 648–57. https://doi.org/10.1046/j.1523-1739.2001.015003648.x.

Proosdij, André S. J. van, Marc S. M. Sosef, Jan J. Wieringa, and Niels Raes. 2016. "Minimum Required Number of Specimen Records to Develop Accurate Species Distribution Models." *Ecography* 39 (6): 542–52. https://doi.org/10.1111/ecog.01509.

Pulliam, H.R. 2000. "On the Relationship between Niche and Distribution." *Ecology Letters* 3 (4): 349–61. https://doi.org/10.1046/j.1461-0248.2000.00143.x.

Reid, Scott. 2014. "Shopping Center Business." *Biggby Coffee*, 2014.

Root, Terry. 1988. "Environmental Factors Associated with Avian Distributional Boundaries." *Journal of Biogeography* 15 (3): 489. https://doi.org/10.2307/2845278.

Sadahiro, Yukio. 2001. "A PDF-Based Analysis of the Spatial Structure of Retailing." *GeoJournal 52: 237–252*, 16.

Sadalla, Edward K., and Stephen G. Magel. 1980. "The Perception of Traversed Distance." *Environment and Behavior* 12 (1): 65–79. https://doi.org/10.1177/0013916580121005.

Sadalla, Edward K., and Lorin J. Staplin. 1980. "The Perception of Traversed Distance: Intersections." *Environment and Behavior* 12 (2): 167–82. https://doi.org/10.1177/0013916580122003.

Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge:

Cambridge University Press. https://doi.org/10.1017/CBO9781107298019.

Sherrouse, Benson C., Jessica M. Clement, and Darius J. Semmens. 2011. “A GIS Application for Assessing, Mapping, and Quantifying the Social Values of Ecosystem Services.” *Applied Geography* 31 (2): 748–60. https://doi.org/10.1016/j.apgeog.2010.08.002.

Shihab, Ibne Farabi, Maliha Moonwara Oishi, Samiul Islam, Kalyan Banik, and Hossain Arif. 2018. “A Machine Learning Approach to Suggest Ideal Geographical Location for New Restaurant Establishment.” In *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 1–5. Malambe, Sri Lanka: IEEE. https://doi.org/10.1109/R10-HTC.2018.8629845.

State of Iowa data portal. 2019. “Iowa Gross Domestic Product by Quarter and Industry.” Data.Iowa.Gov. 2019. https://data.iowa.gov/Economic-Statistics/Iowa-Gross-Domestic-Product-by-Quarter-and-Industr/f2xe-wr7z/data.

Sugumaran, Ramanathan, and John Degroote. 2011. “Spatial Decision Support Systems,” 510.

Tam, Kar Yan, and Melody Y. Kiang. 1992. “Managerial Applications of Neural Networks: The Case of Bank Failure Predictions.” *Management Science* 38 (7): 926–47. https://doi.org/10.1287/mnsc.38.7.926.

Tzeng, Gwo-Hshiung, Mei-Hwa Teng, June-Jye Chen, and Serafim Opricovic. 2002. “Multicriteria Selection for a Restaurant Location in Taipei.” *International Journal of Hospitality Management* 21 (2): 171–87. https://doi.org/10.1016/S0278-4319(02)00005-1.

Vallat, Raphael. 2018. “Pingouin: Statistics in Python.” *Journal of Open Source Software* 3 (31): 1026. https://doi.org/10.21105/joss.01026.

Ward, Kevin. 2005. “Entrepreneurial Urbanism and the Management of the Contemporary City: The Example of Business Improvement Districts.” School of Geography University of Manchester.

Witlox, F. 2003. “MATISSE: A Relational Expert System for Industrial Site Selection.” *Expert Systems with Applications* 24 (1): 133–44. https://doi.org/10.1016/S0957-4174(02)00091-X.

Yang, Yang, Jingyin Tang, Hao Luo, and Rob Law. 2015. “Hotel Location Evaluation: A Combination of Machine Learning Tools and Web GIS.” *International Journal of Hospitality Management* 47 (May): 14–24. https://doi.org/10.1016/j.ijhm.2015.02.008.

Yeh, Anthony G-O. 1999. “Urban Planning and GIS.” *Geographical Information Systems* 2, no. 877-888 (1999): 1.

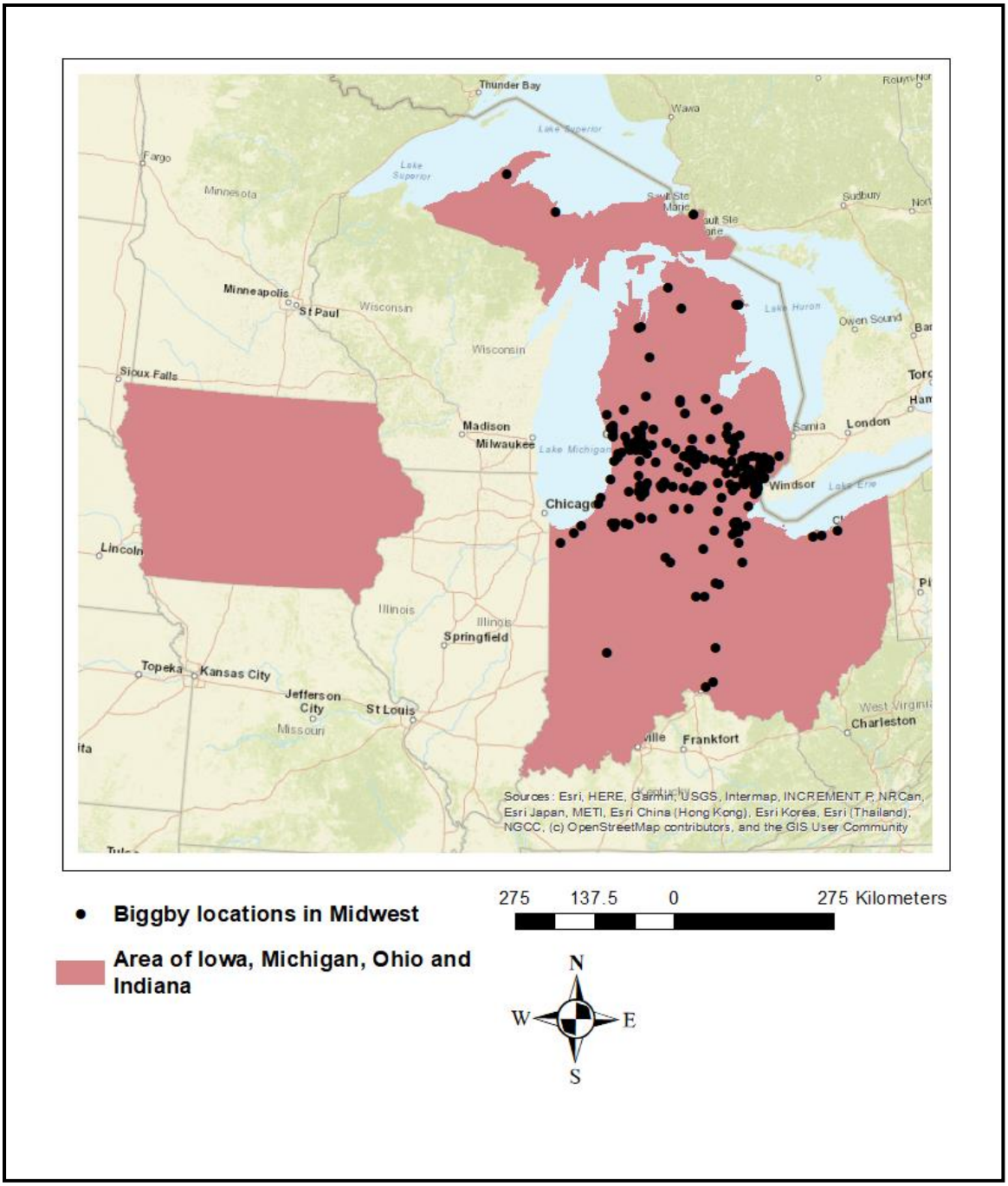
Yıldız, Nurdan, and Fatih Tüysüz. 2018. “A Hybrid Multi-Criteria Decision Making Approach for Strategic Retail Location Investment: Application to Turkish Food Retailing.” *Socio-Economic Planning Sciences*, March, 100619. https://doi.org/10.1016/j.seps.2018.02.006.

Young, Nick, Lane Carter, and Paul Evangelista. 2011. “A MaxEnt Model v3.3.3e Tutorial (ArcGIS V10).”

APPENDIX

Model Inputs

Here are all the variables materials that were used in models in this research. Maps display how they are distribute in space. All variables were displayed in the same style, in order to make it easier to understand in what format they were used in the algorithms.



.

Figure 23. Model input: locations of Biggby cafes in the Midwest

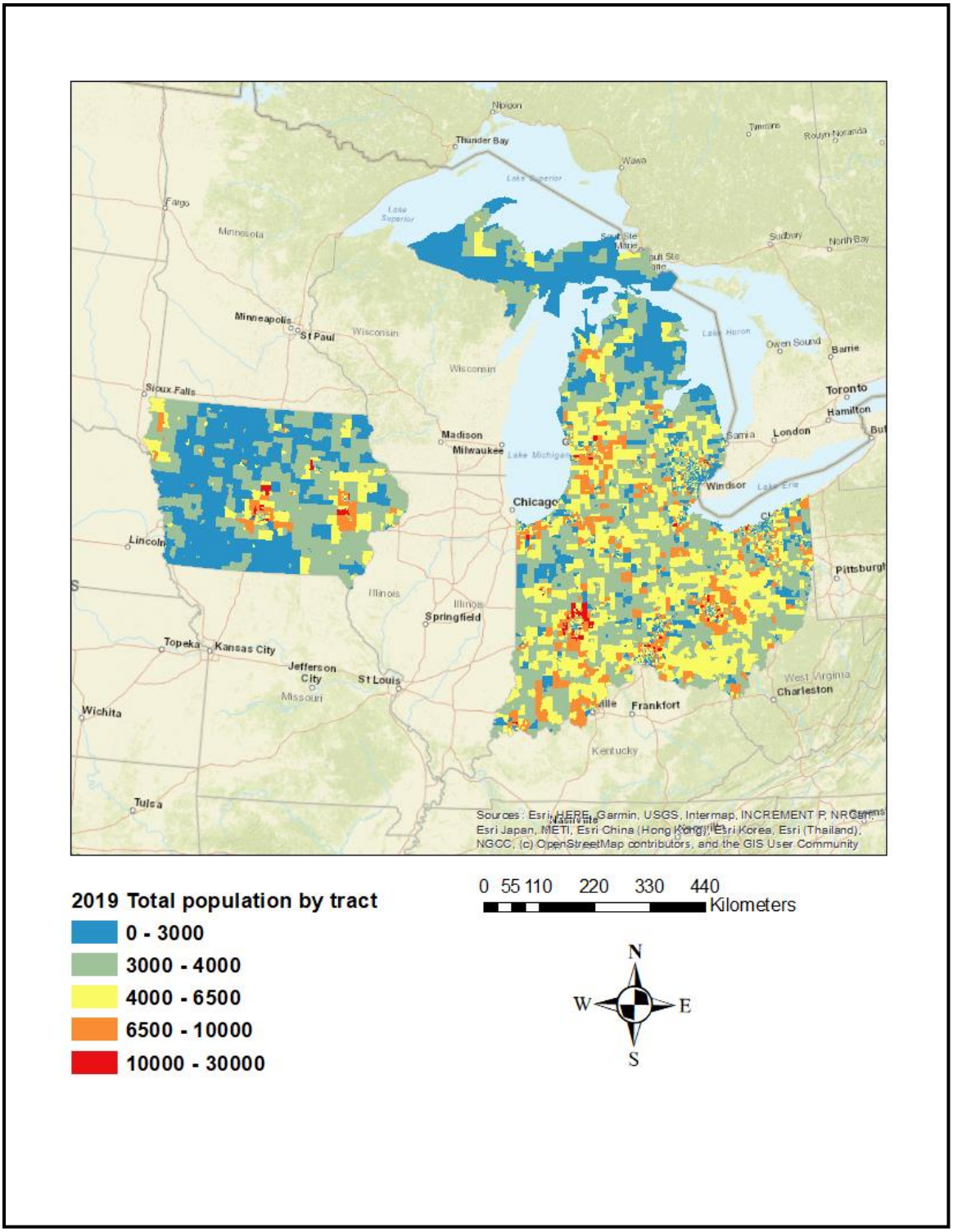


Figure 24. Model input: total population by tract

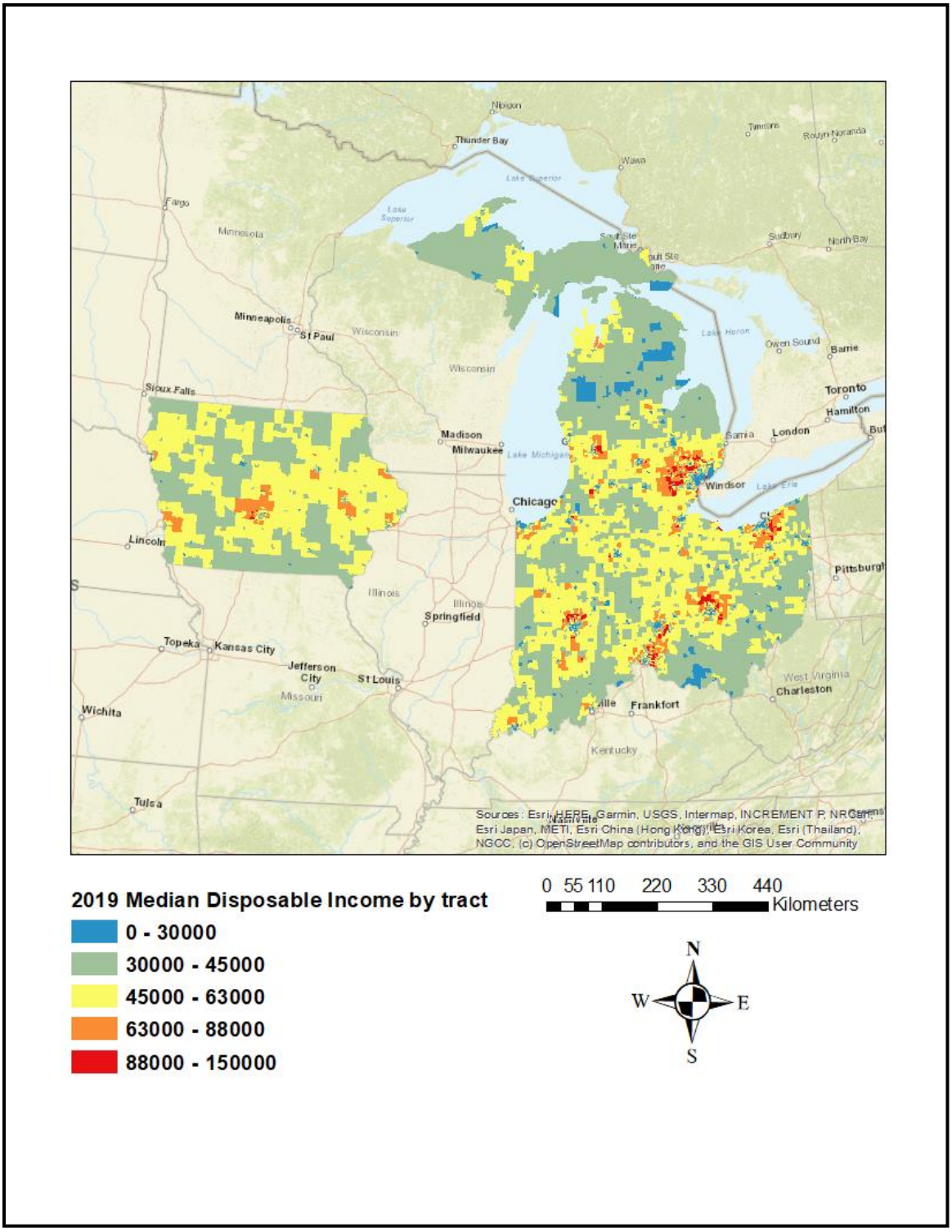


Figure 25. . *Model input:* median disposable income by tract

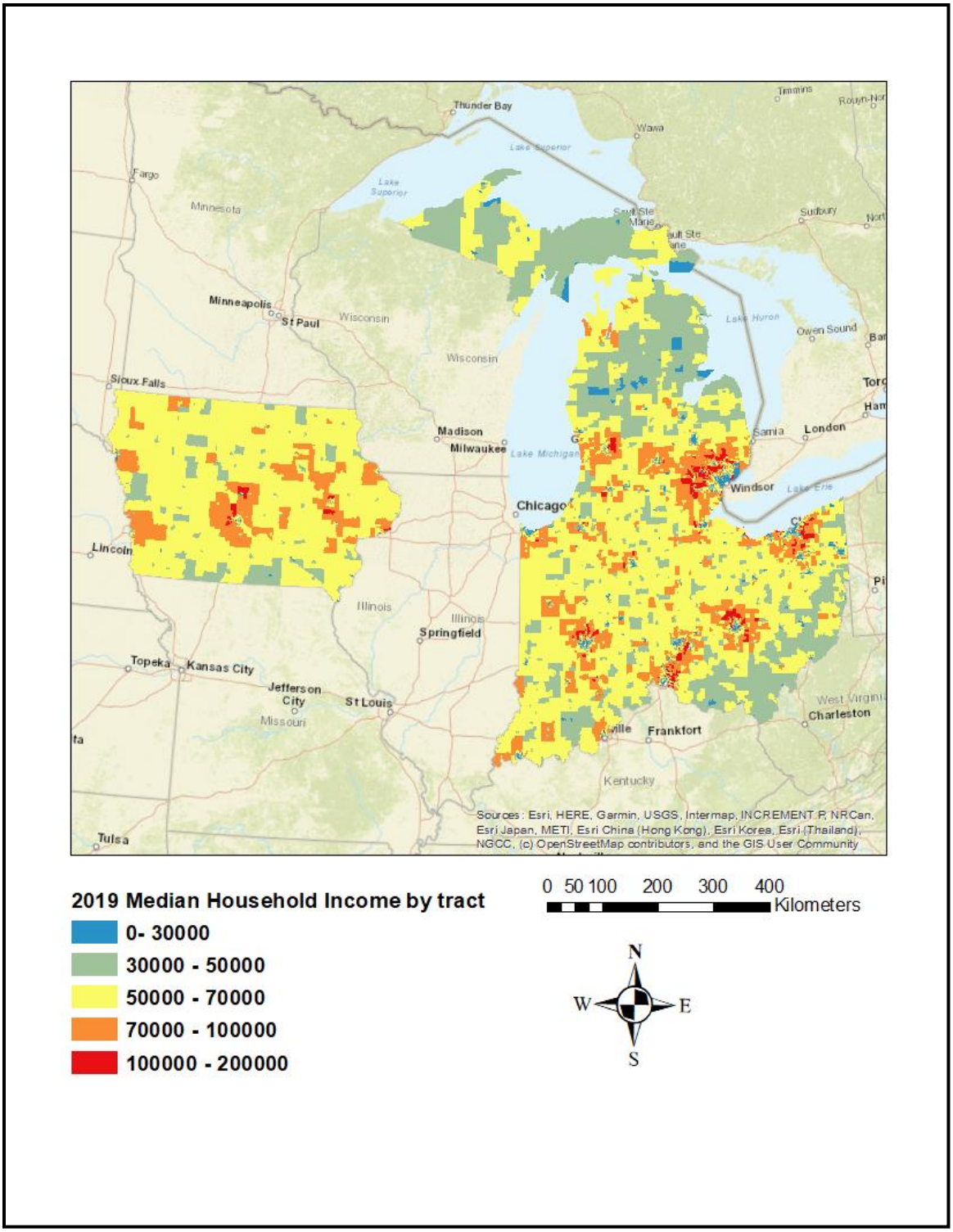


Figure 26. Model input: median household income by tract

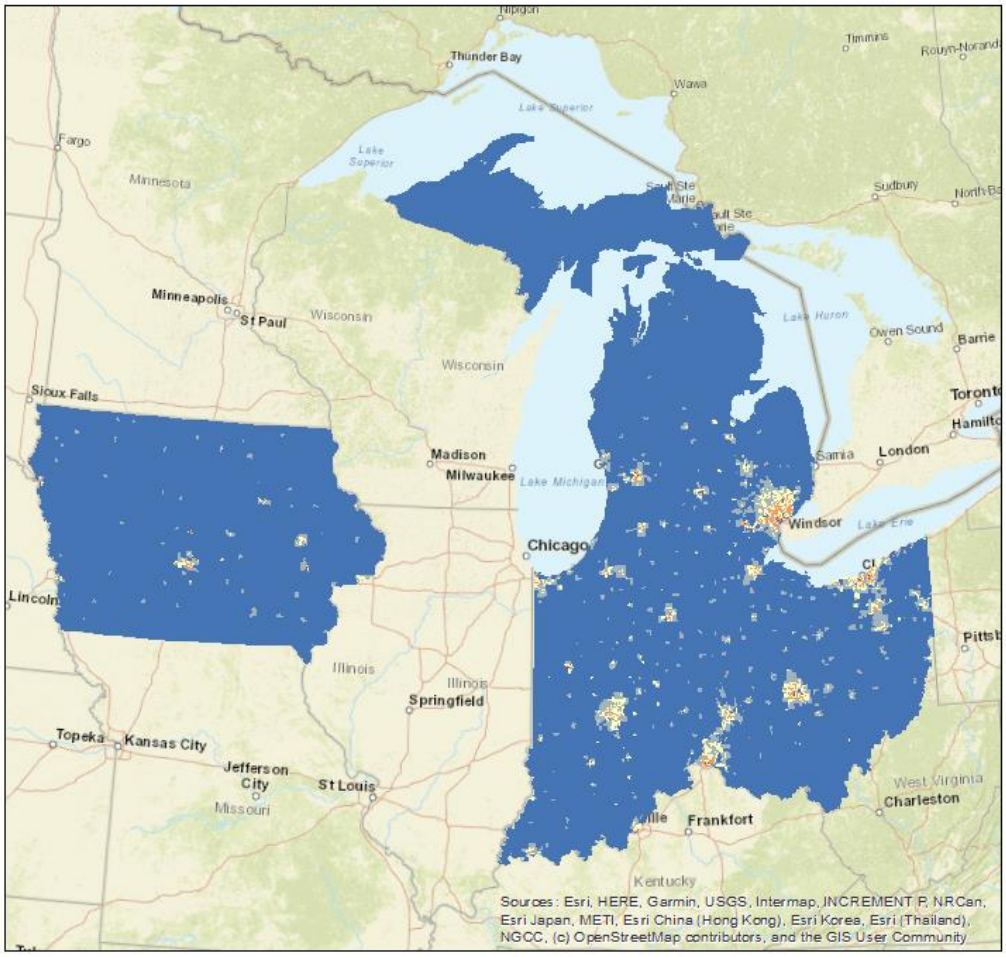


Figure 27. . Model input: population density by tract

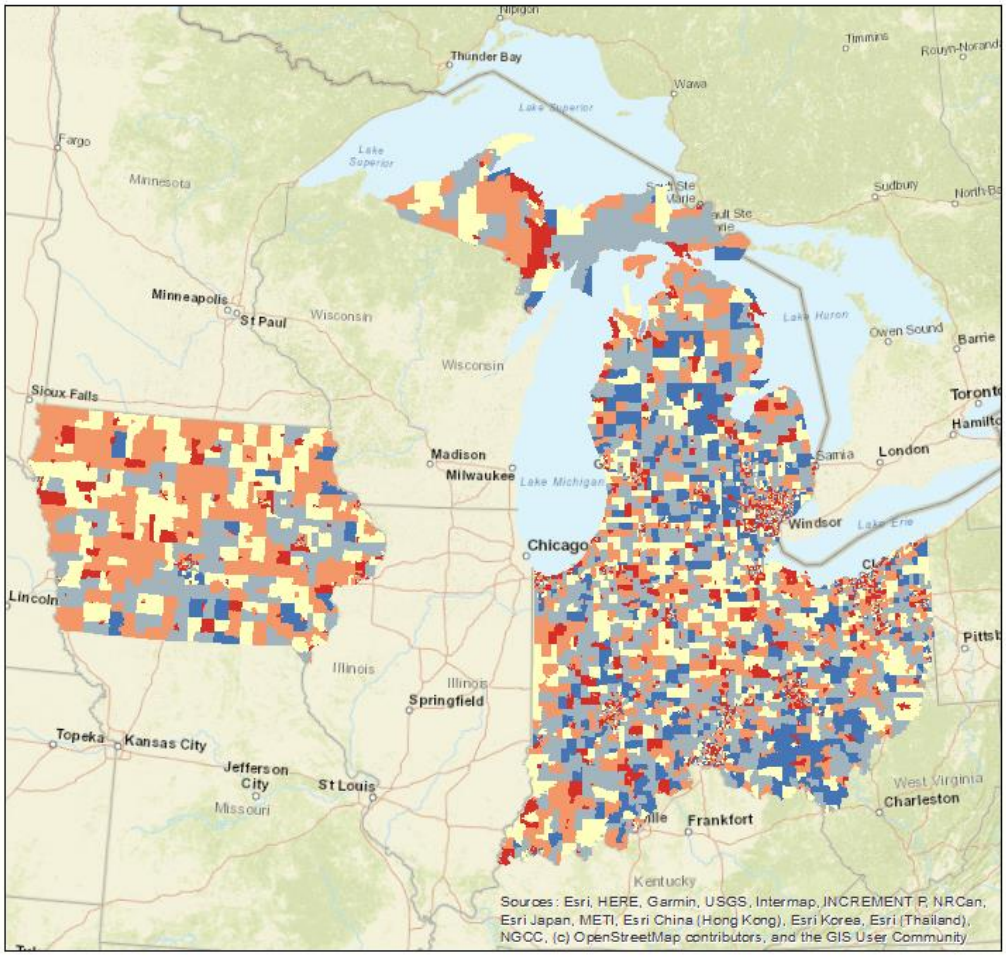


Figure 28. . Model input: percent of working population by tract

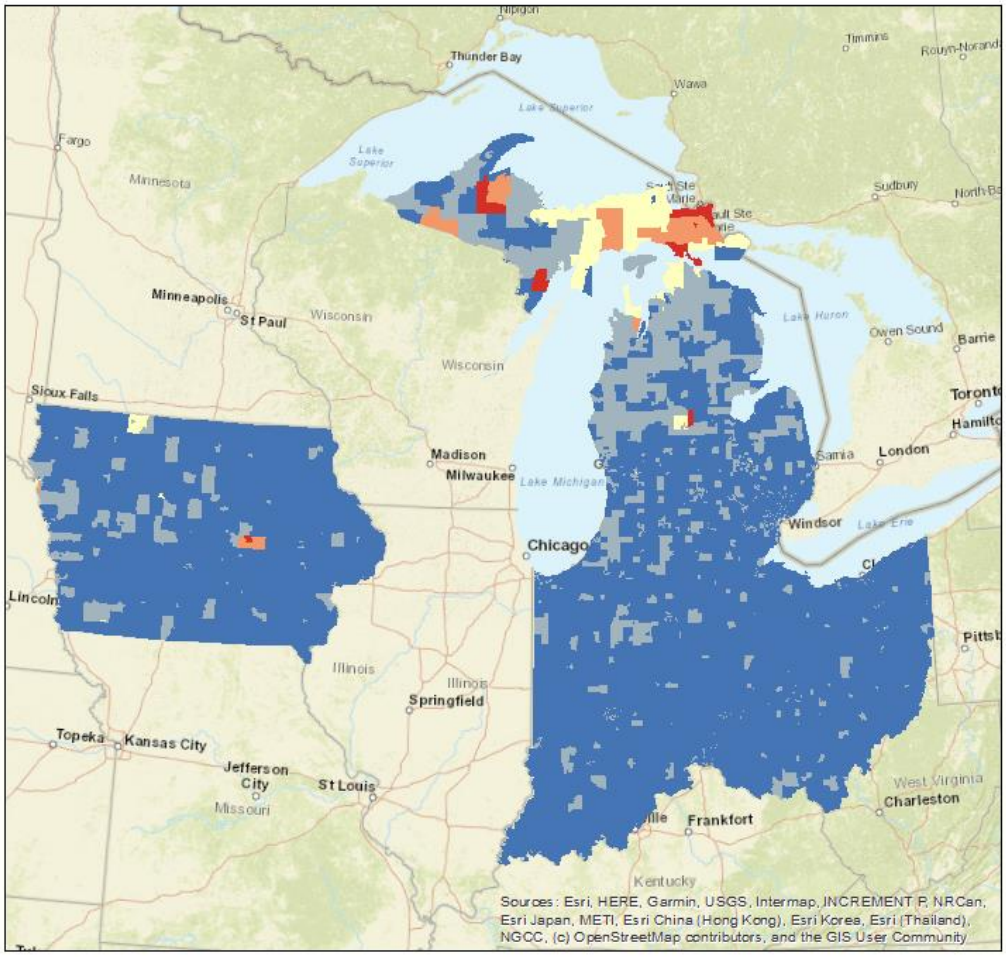


Figure 29. . Model input: percent of American Indigenous population by tract

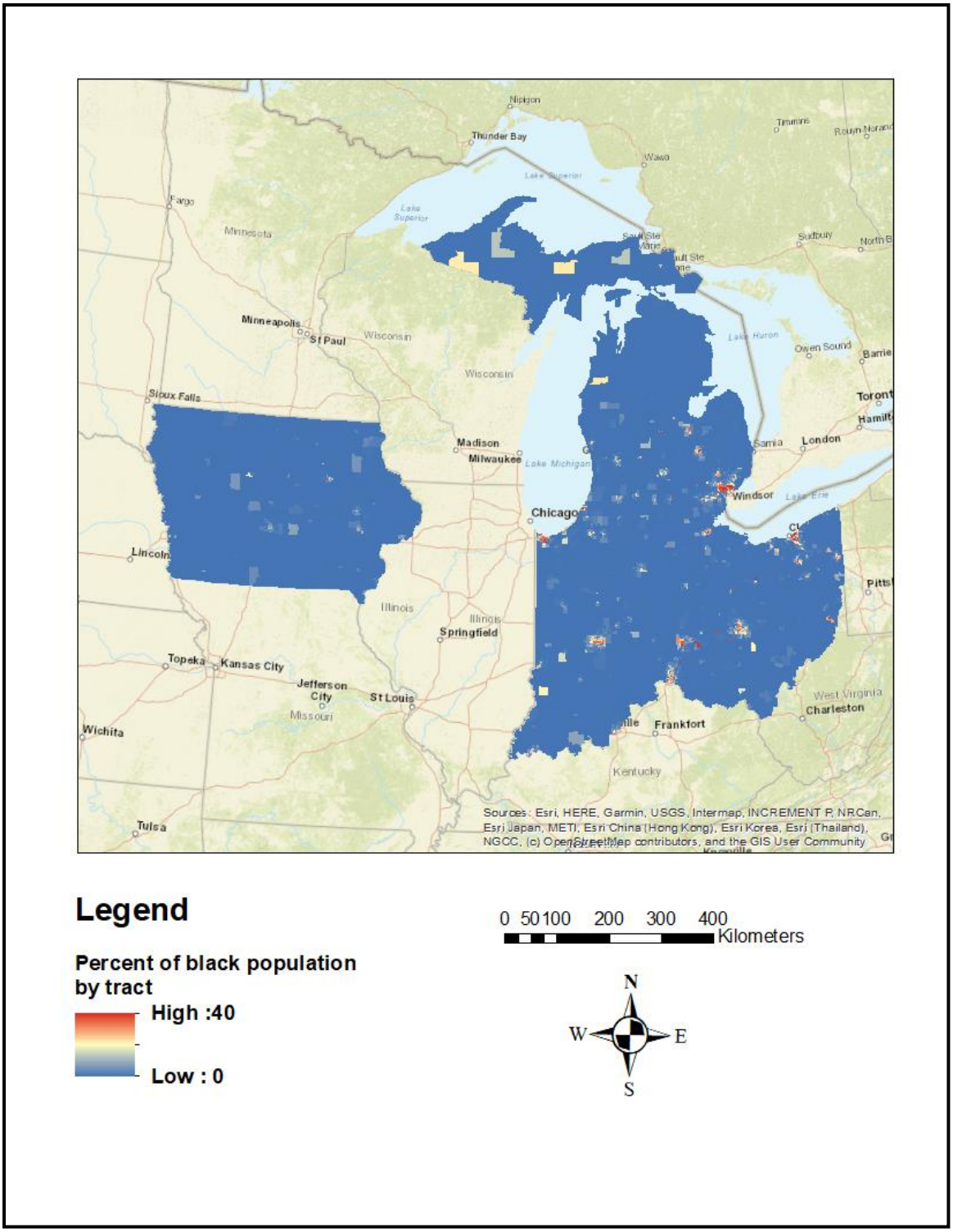


Figure 30. . Model input: percent of American Black population by tract

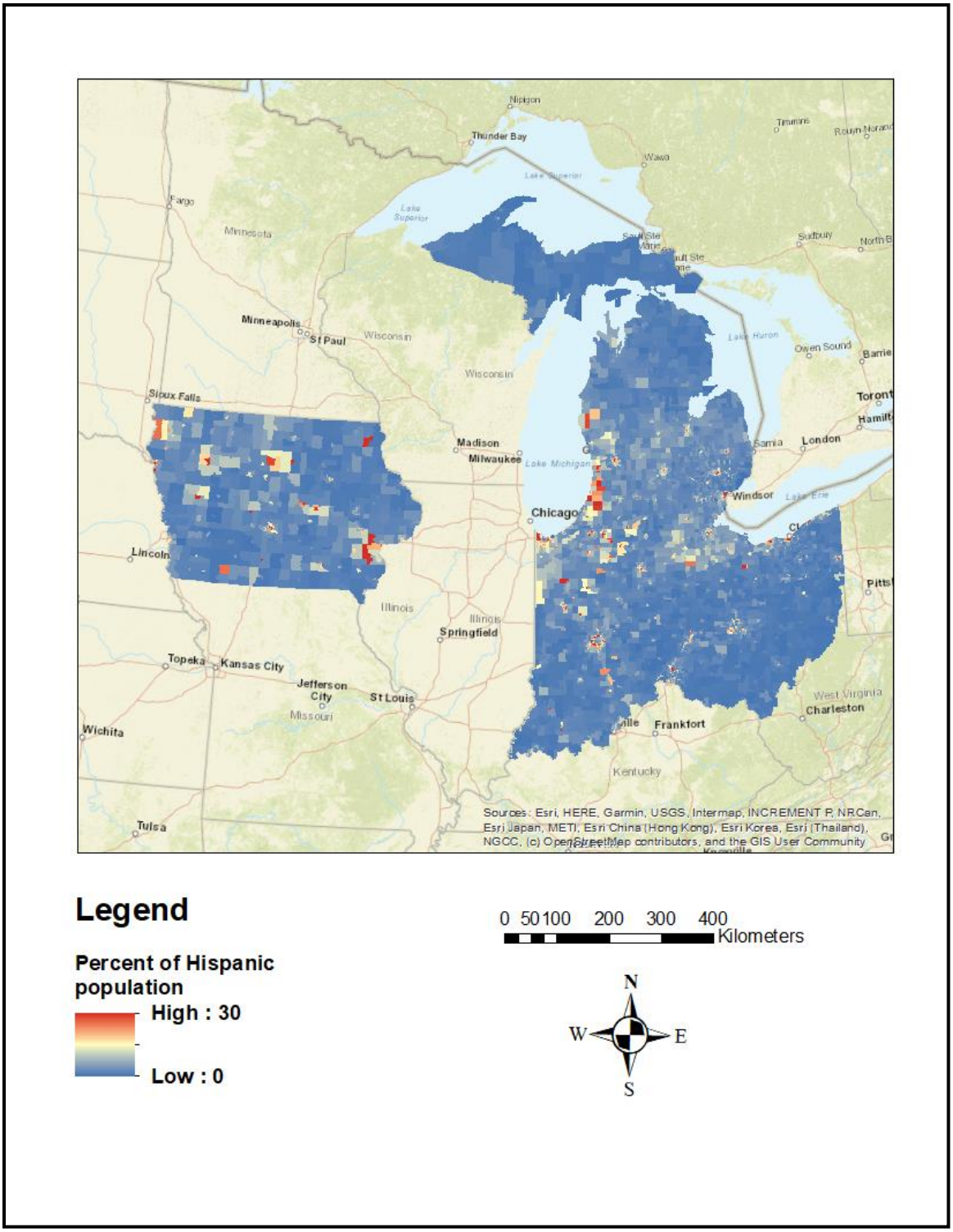


Figure 31. . Model input: percent of Hispanic population by tract

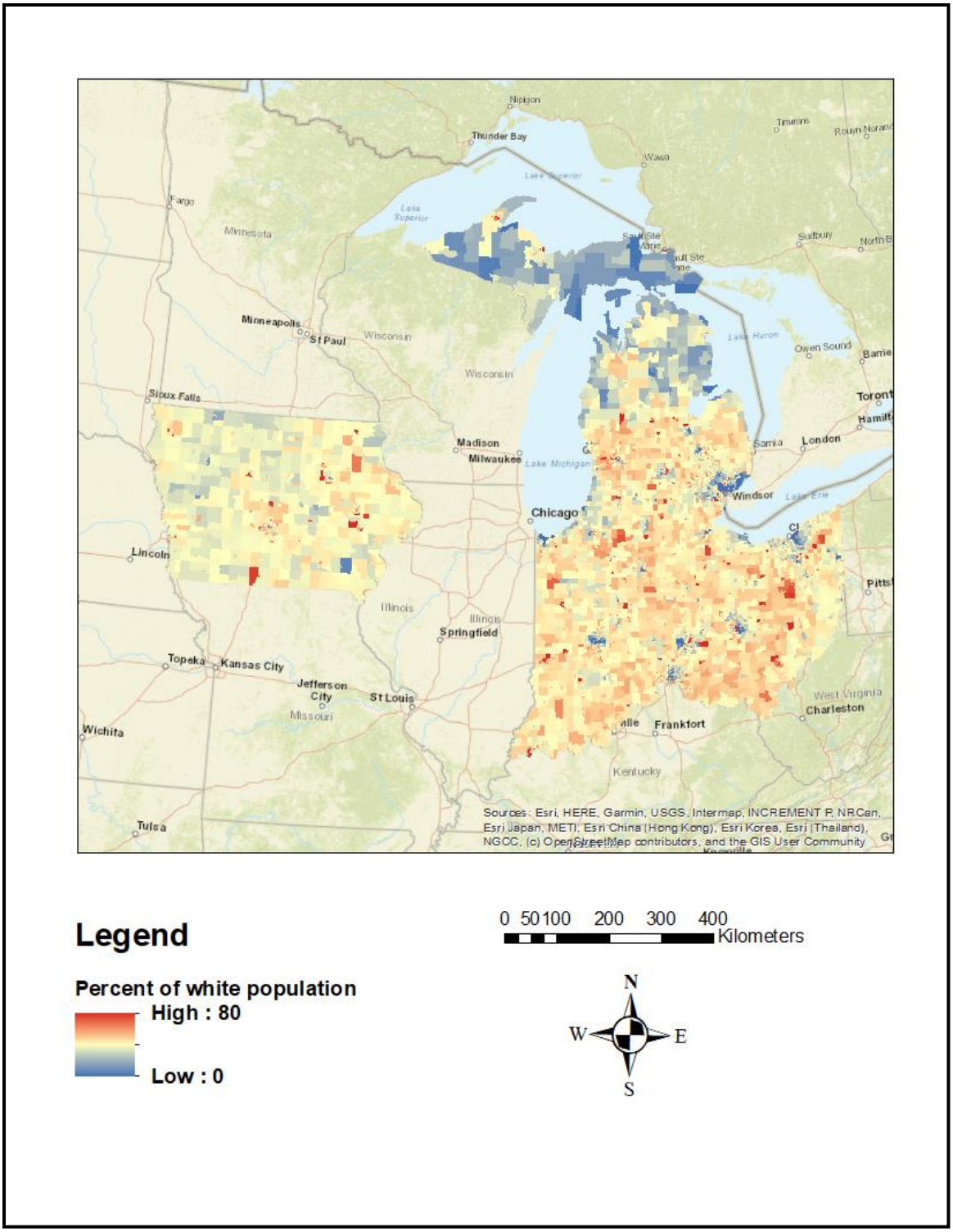


Figure 32. . Model input: percent of White population by tract

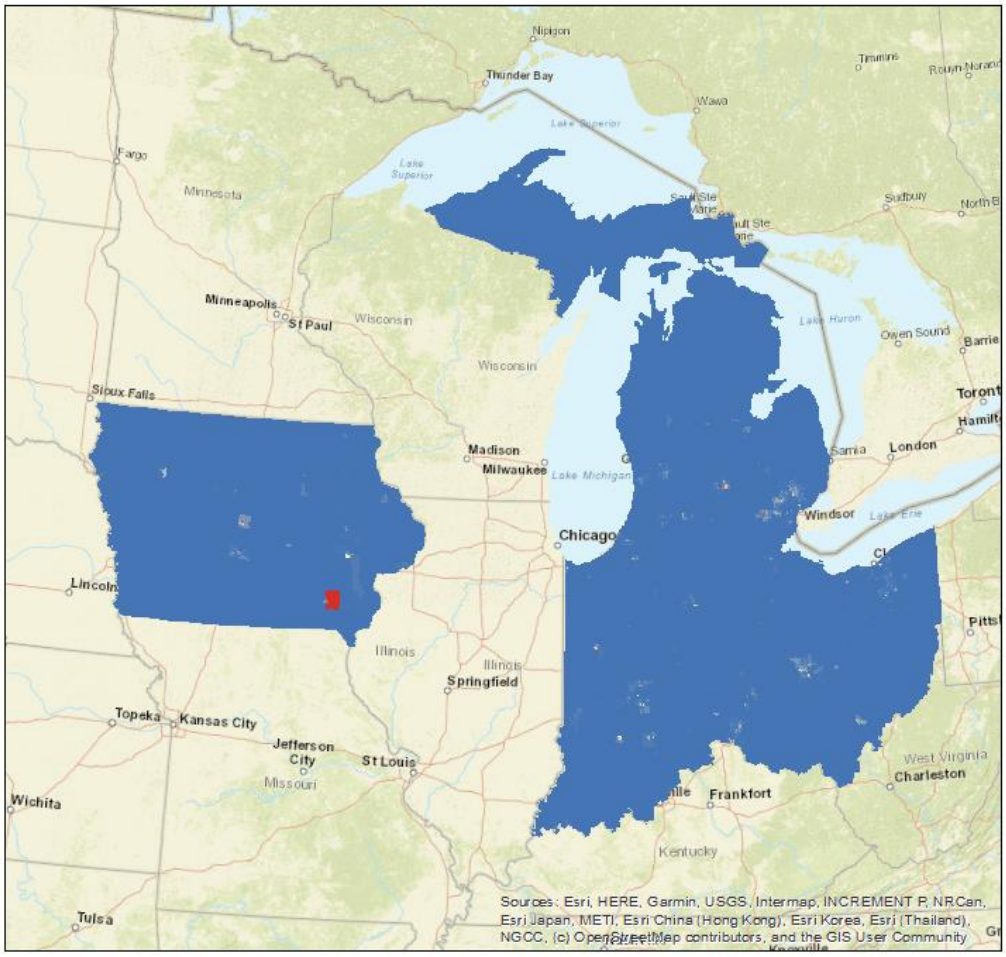


Figure 33. . Model input: percent of Asian people by tract

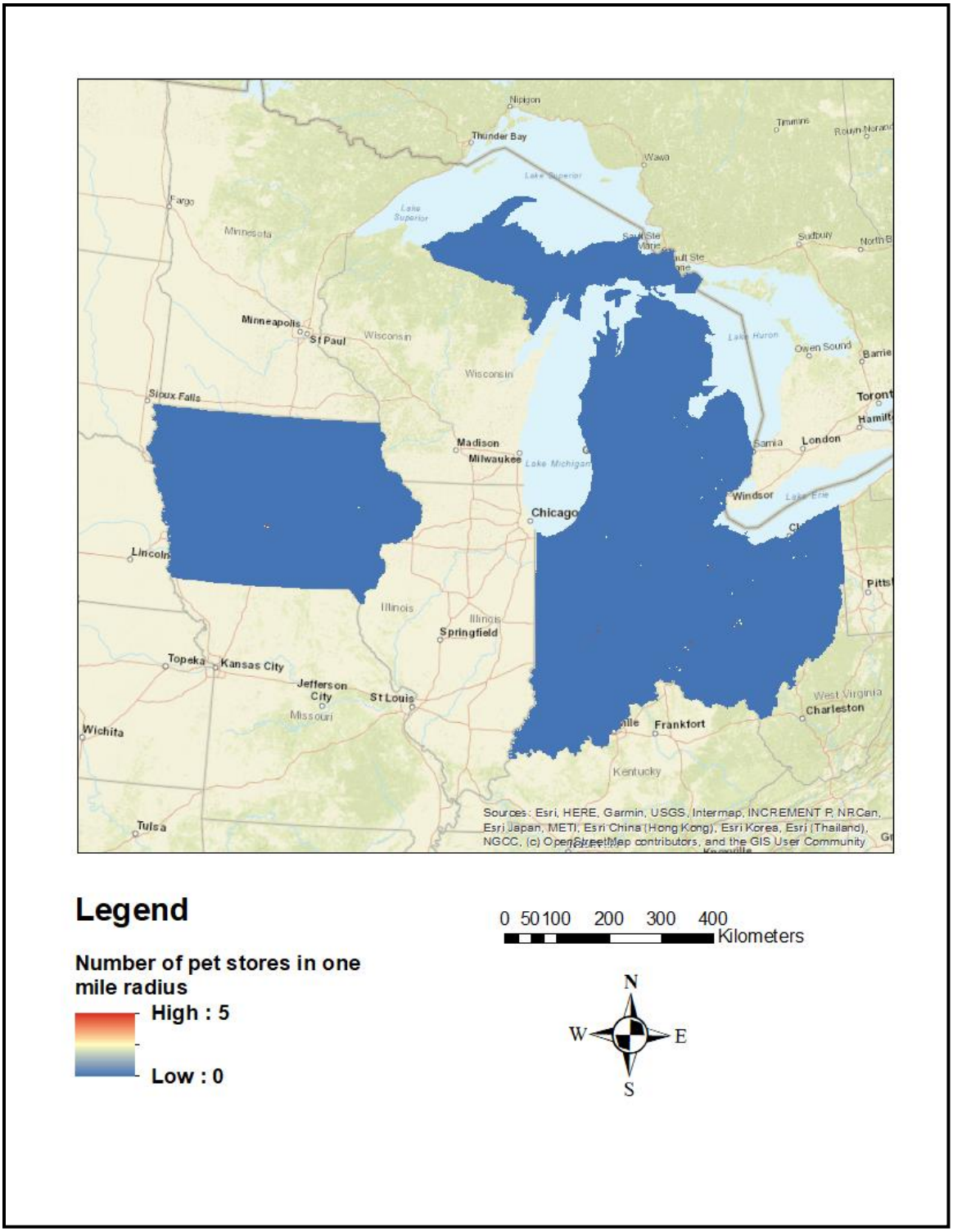


Figure 34. . Model input: number of pet stores in one mile radius

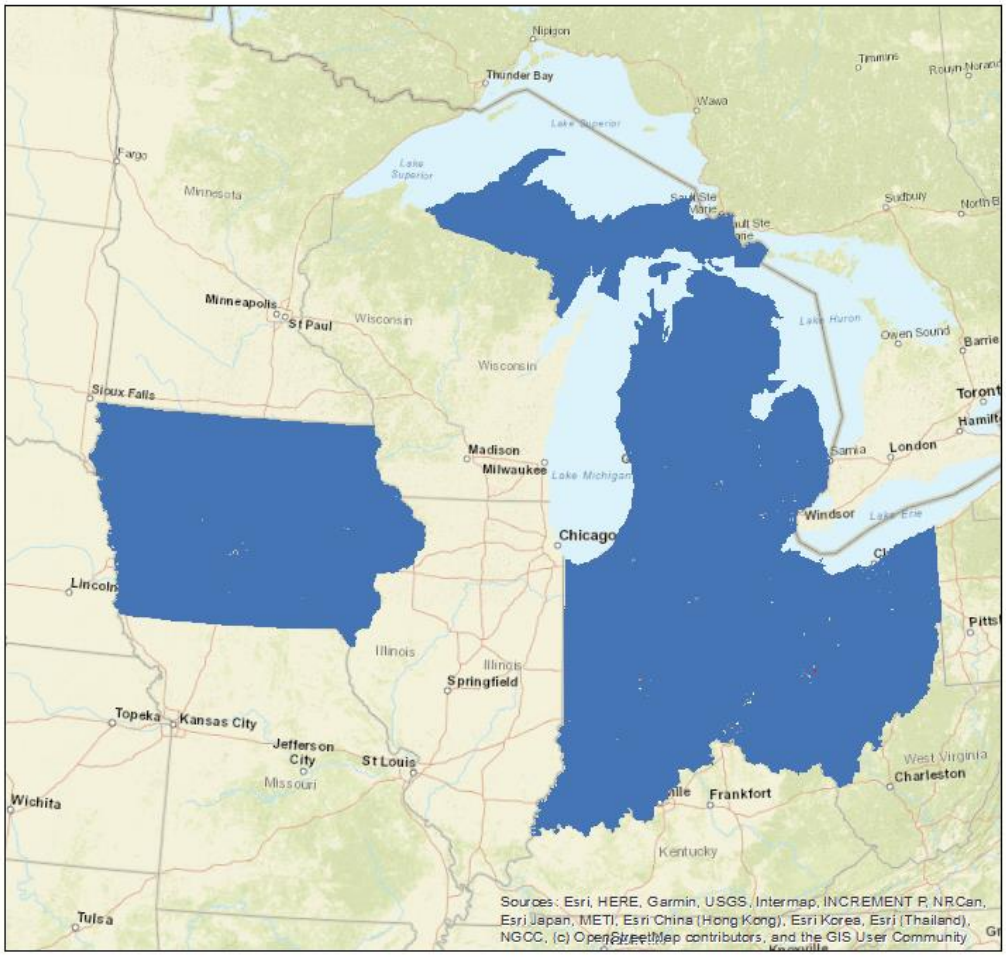


Figure 35. . Model input: number of pharmacies in one mile radius

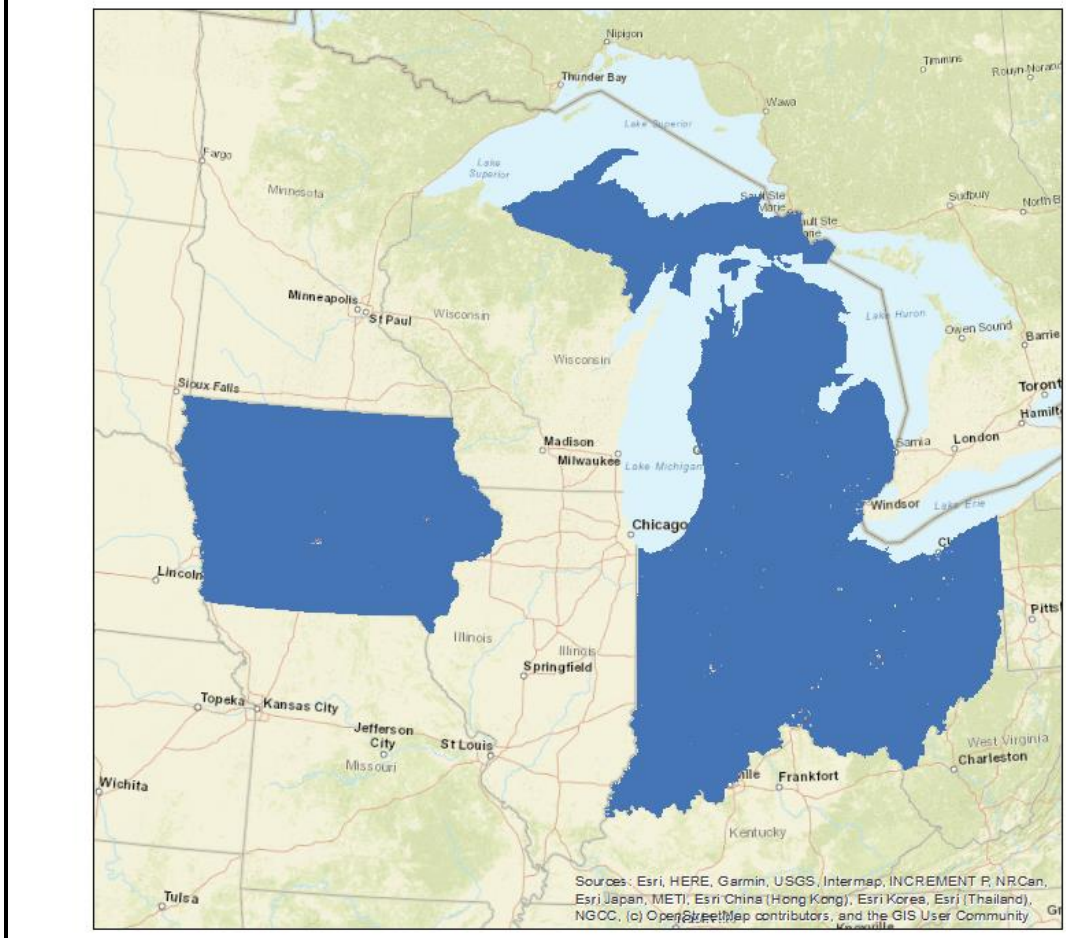


Figure 36. . Model input: number of furniture stores in one mile radius

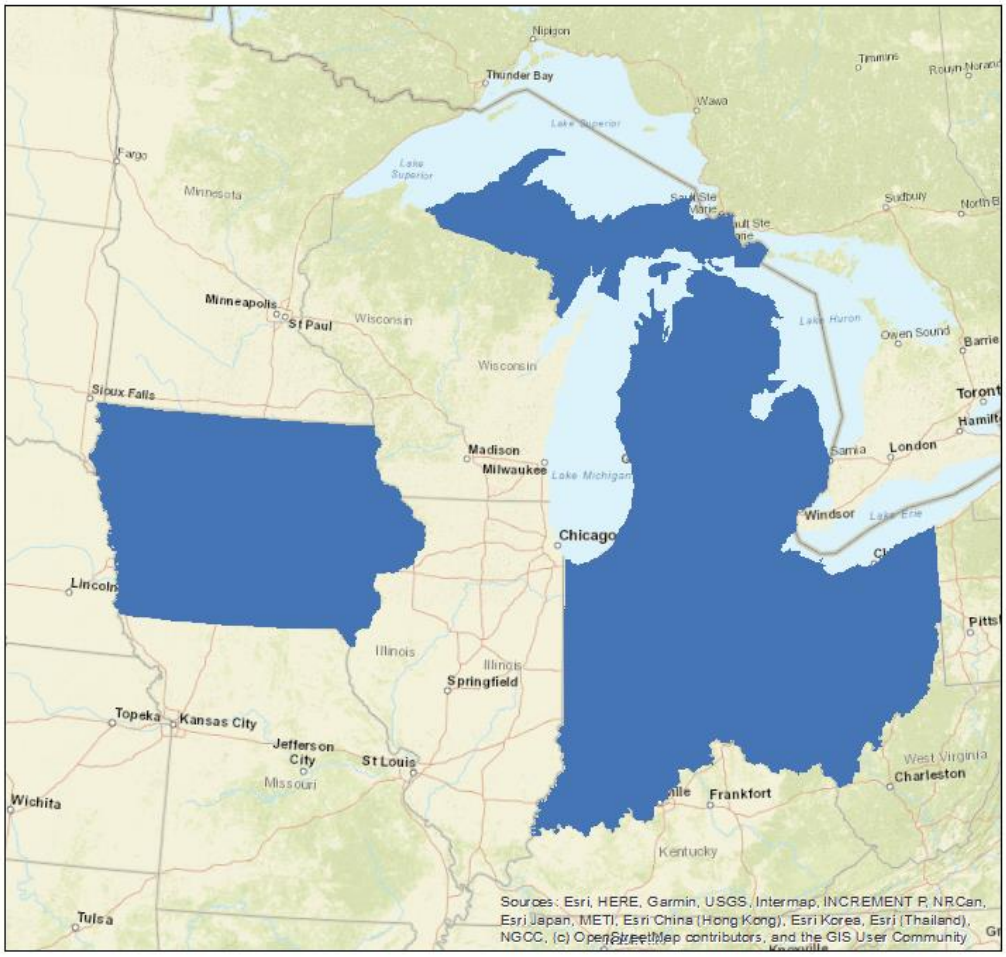


Figure 37. . Model input: number of gas stations in one mile radius

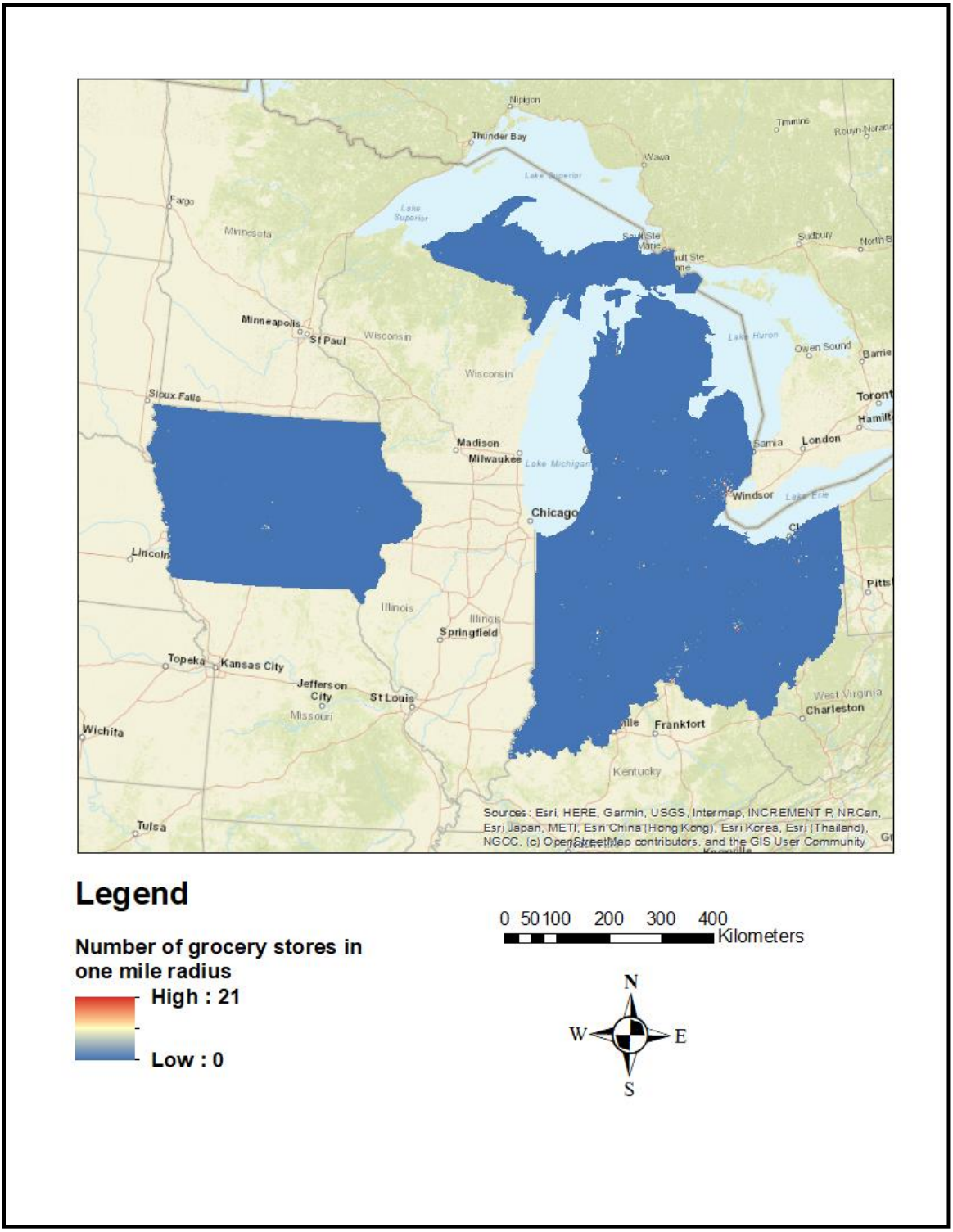


Figure 38. . Model input: number of grocery stores in one mile radius

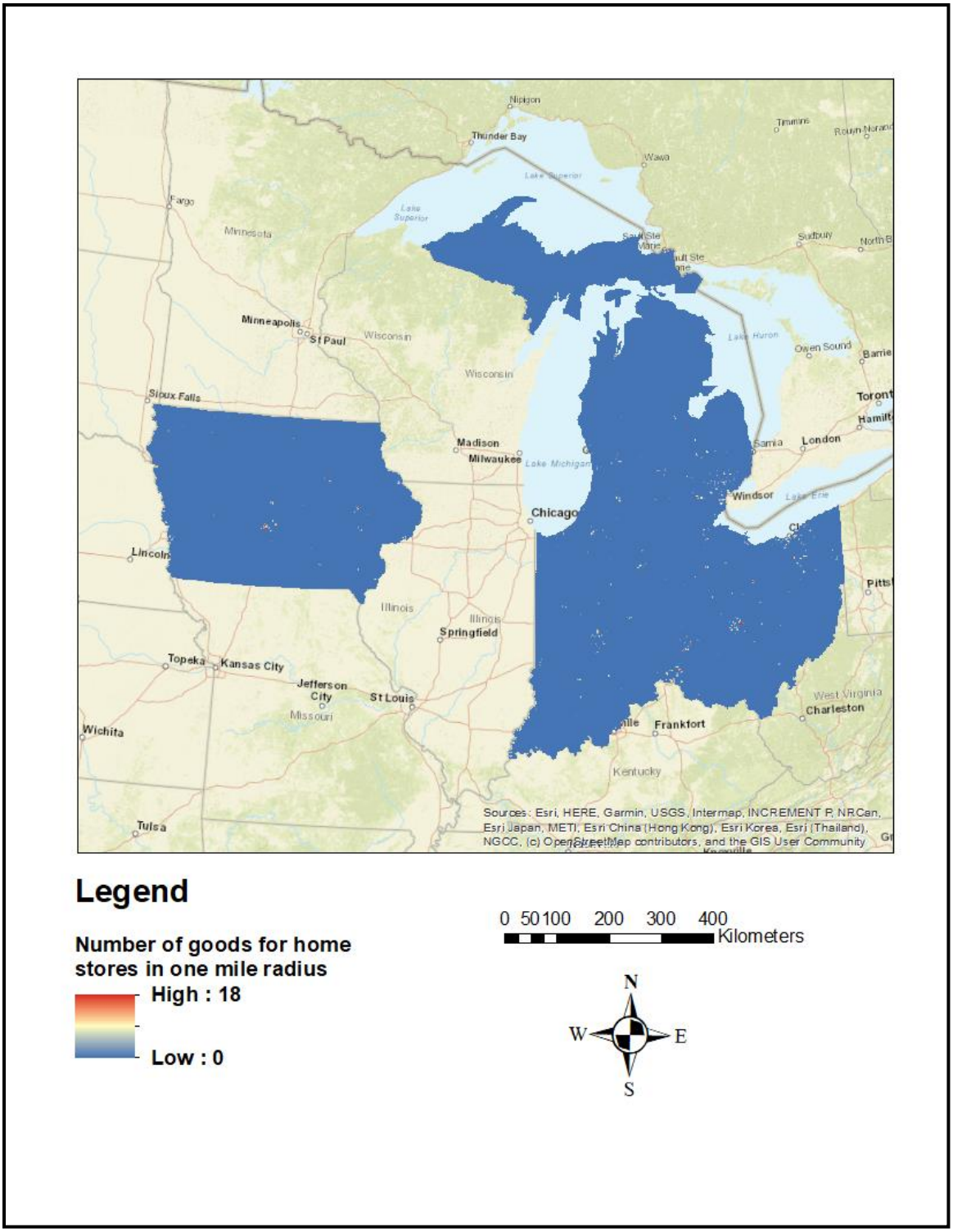


Figure 39. . Model input: number of goods for home stores in one mile radius

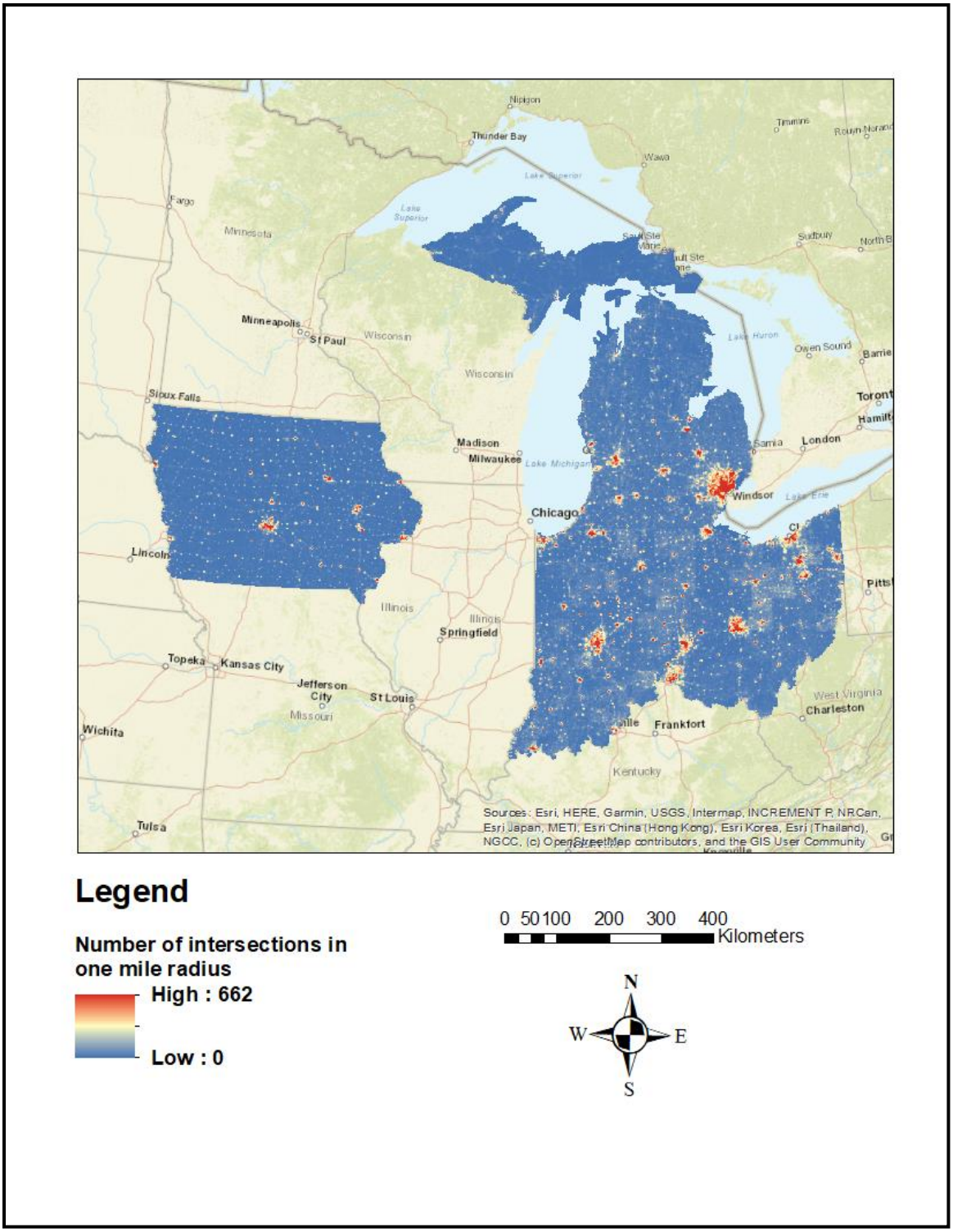


Figure 40. . Model input: number of intersections in one mile radius

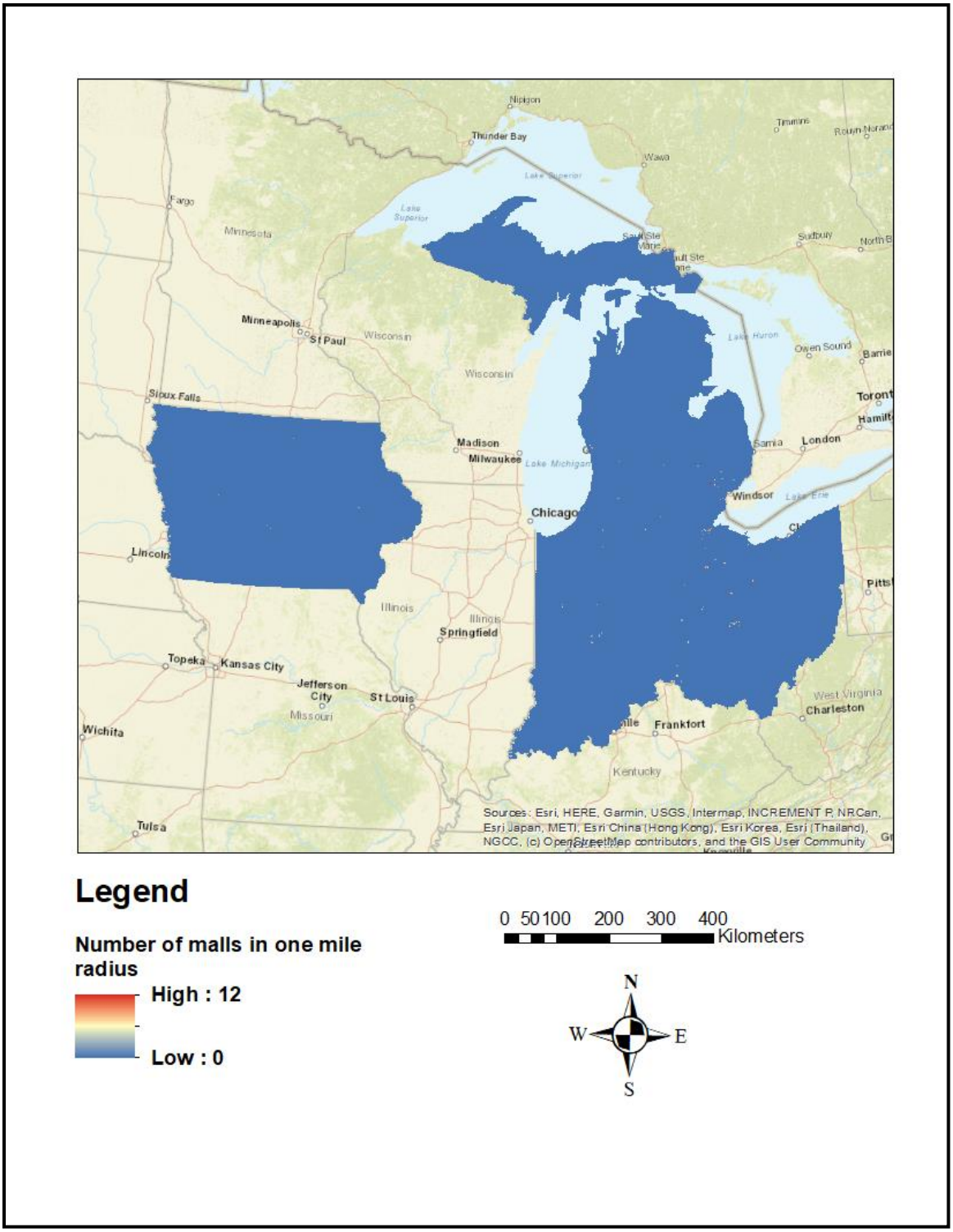


Figure 41. Model input: number of malls in one mile radius

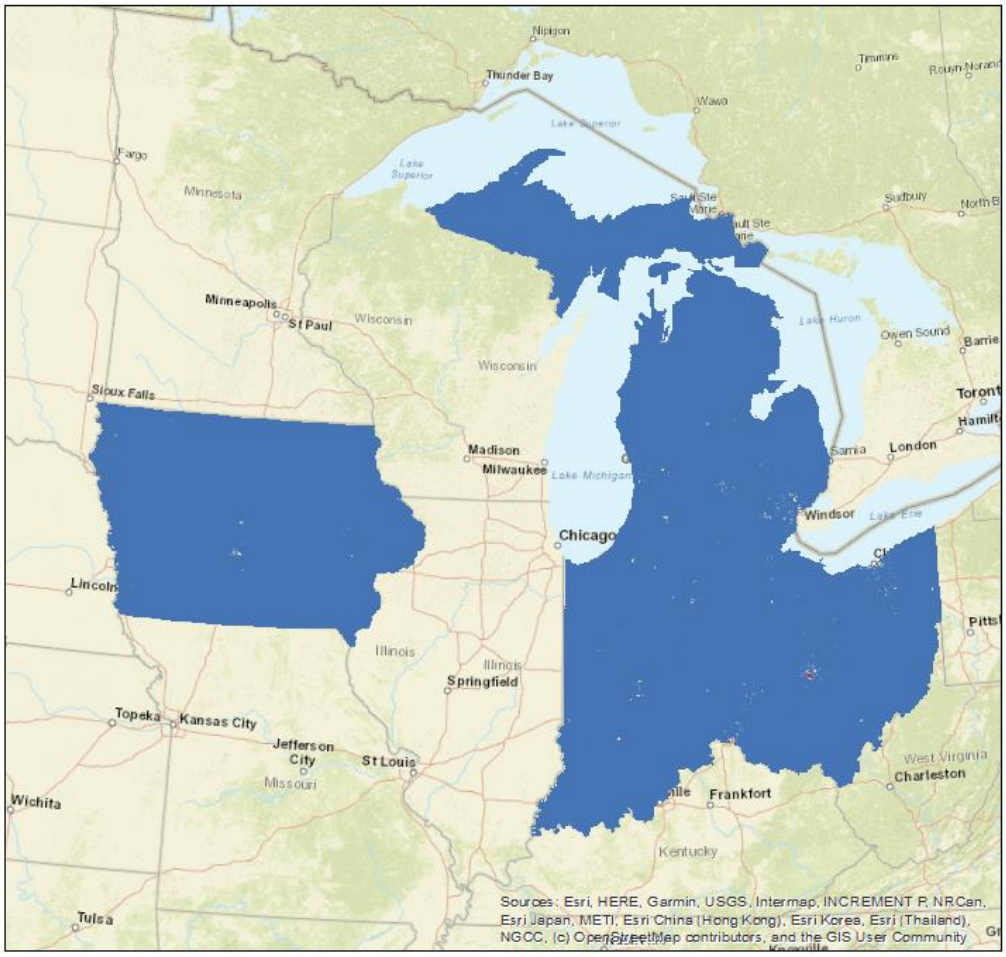


Figure 42. Model input: number of cafes in one mile radius

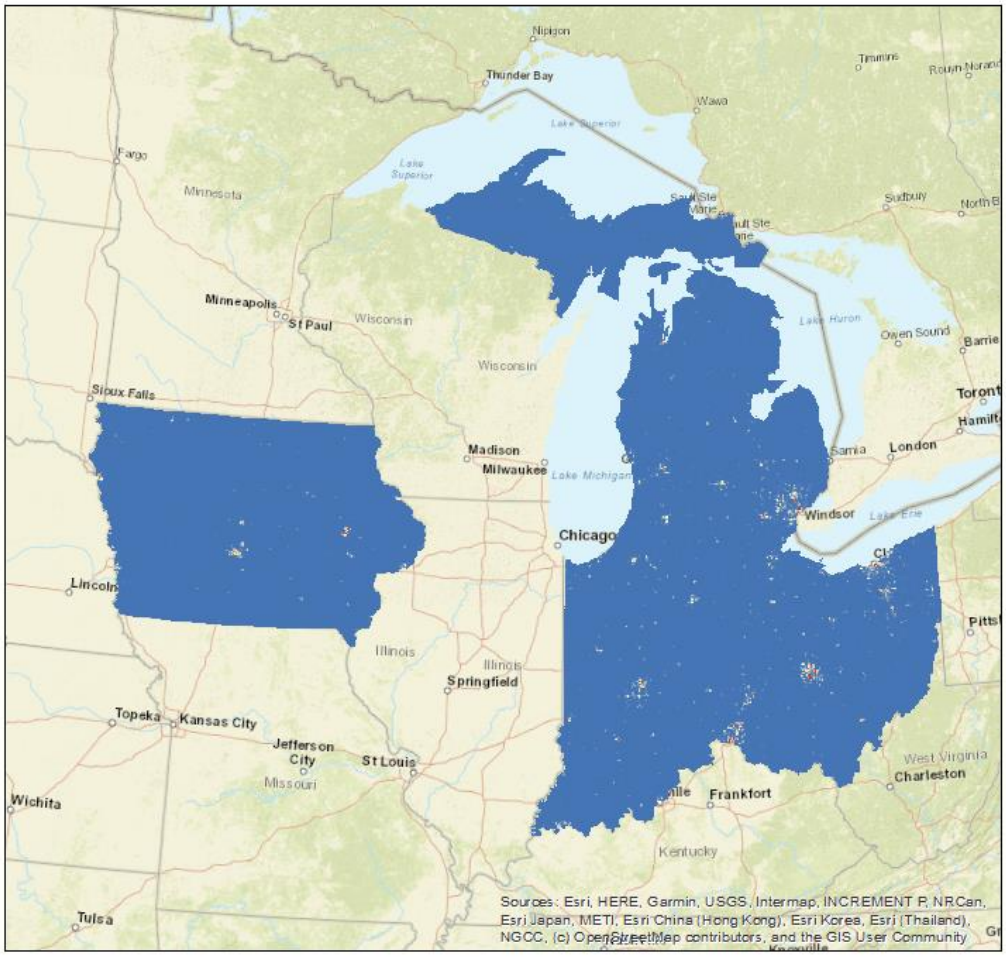


Figure 43. Model input: number of restaurants in one mile radius

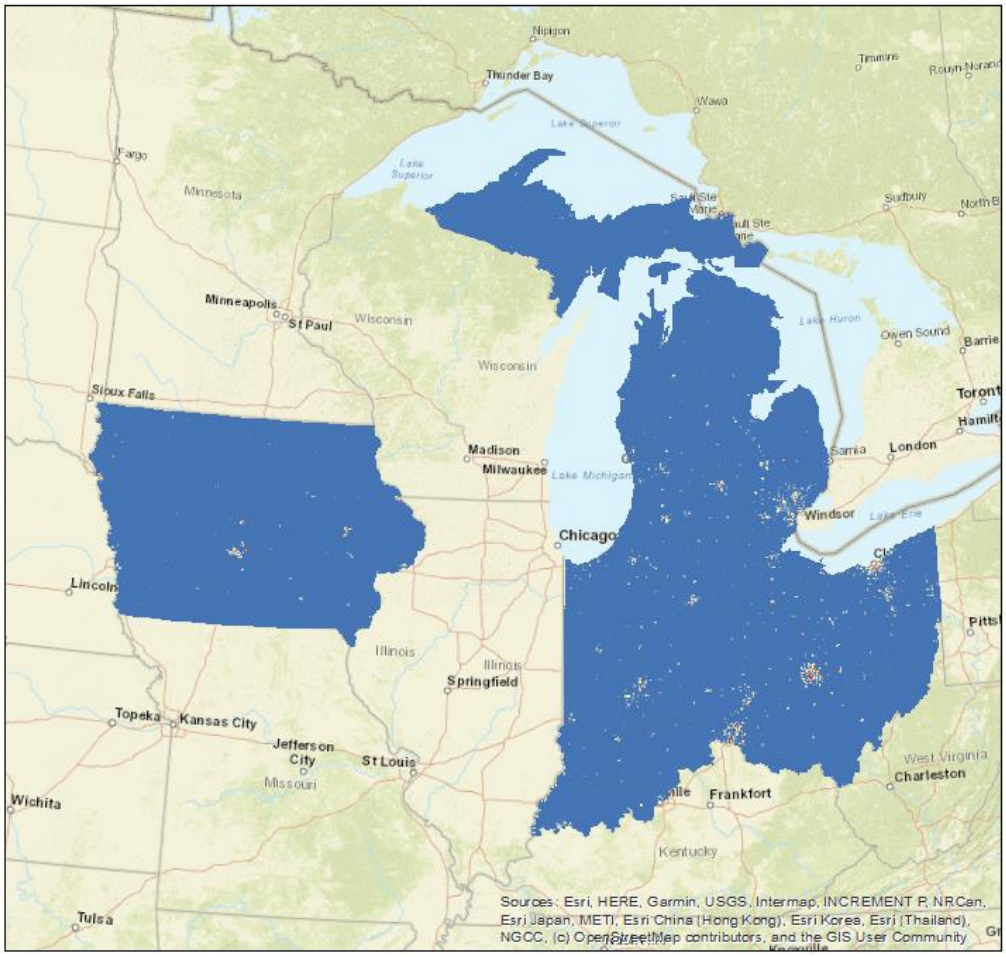


Figure 44. Model input: number of fast food in one mile radius

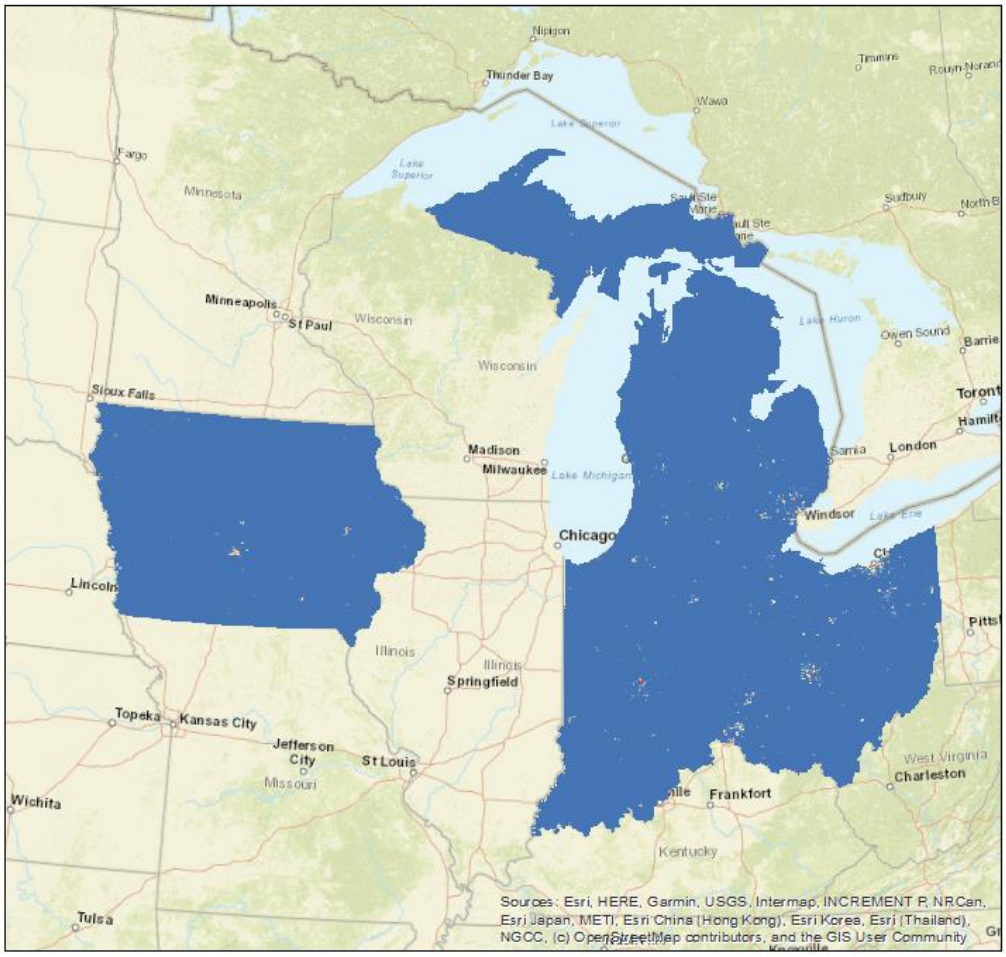


Figure 45. Model input: number of services in one mile radius

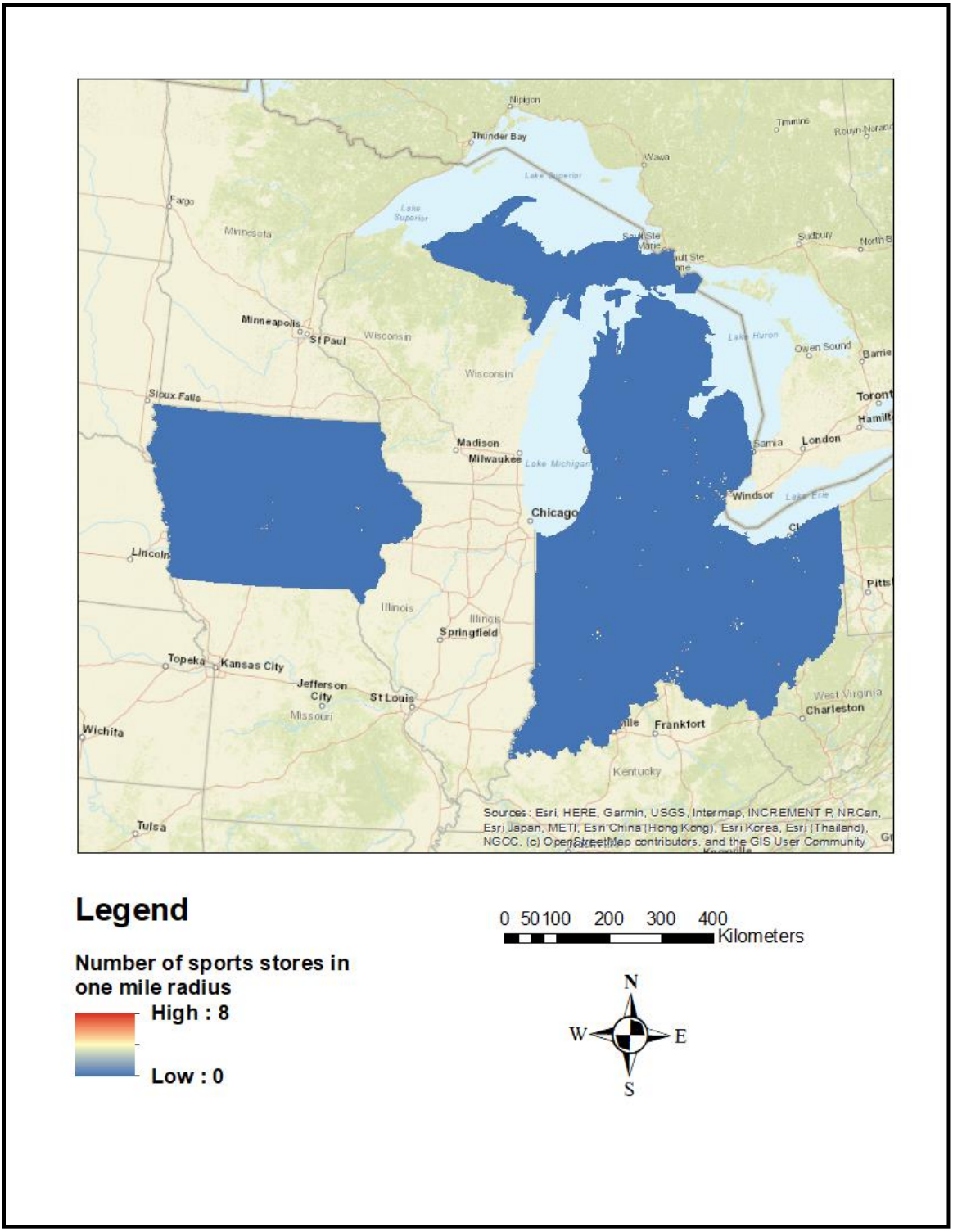


Figure 46. Model input: number of sport stores in one mile radius

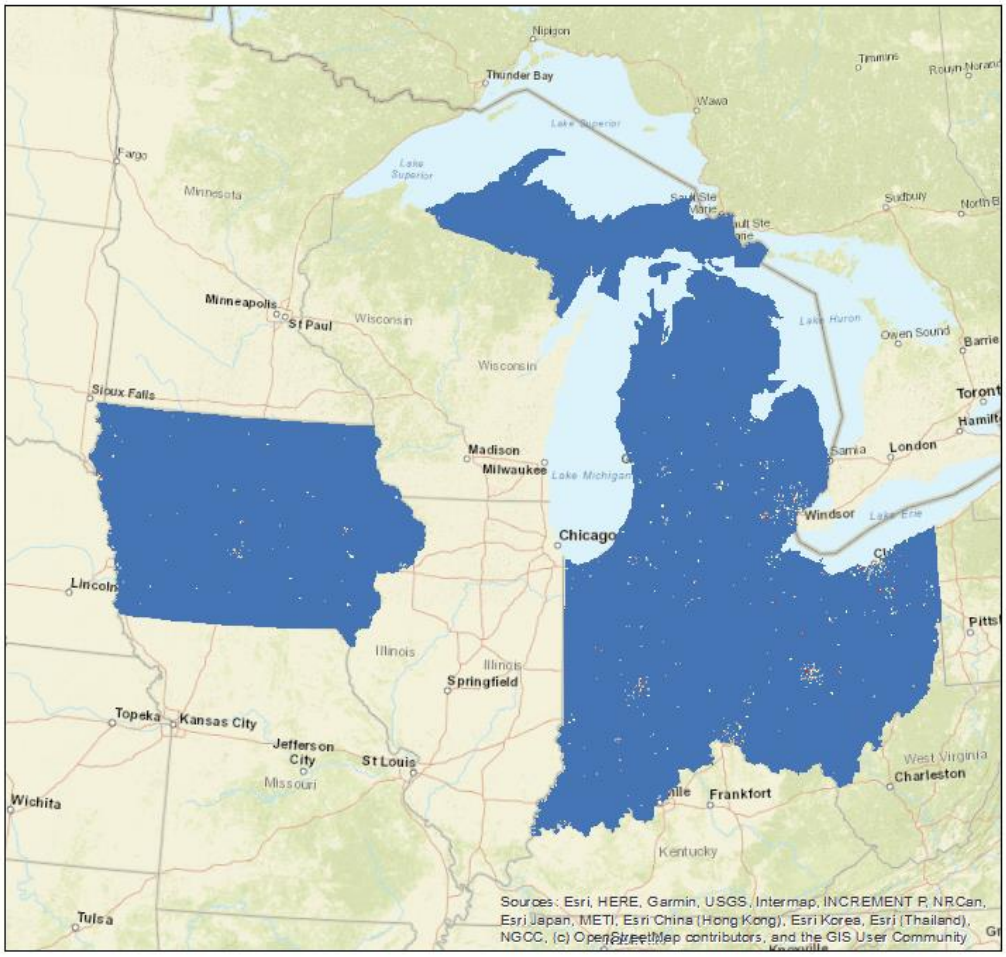


Figure 47. Model input: number of supermarkets in one mile radius

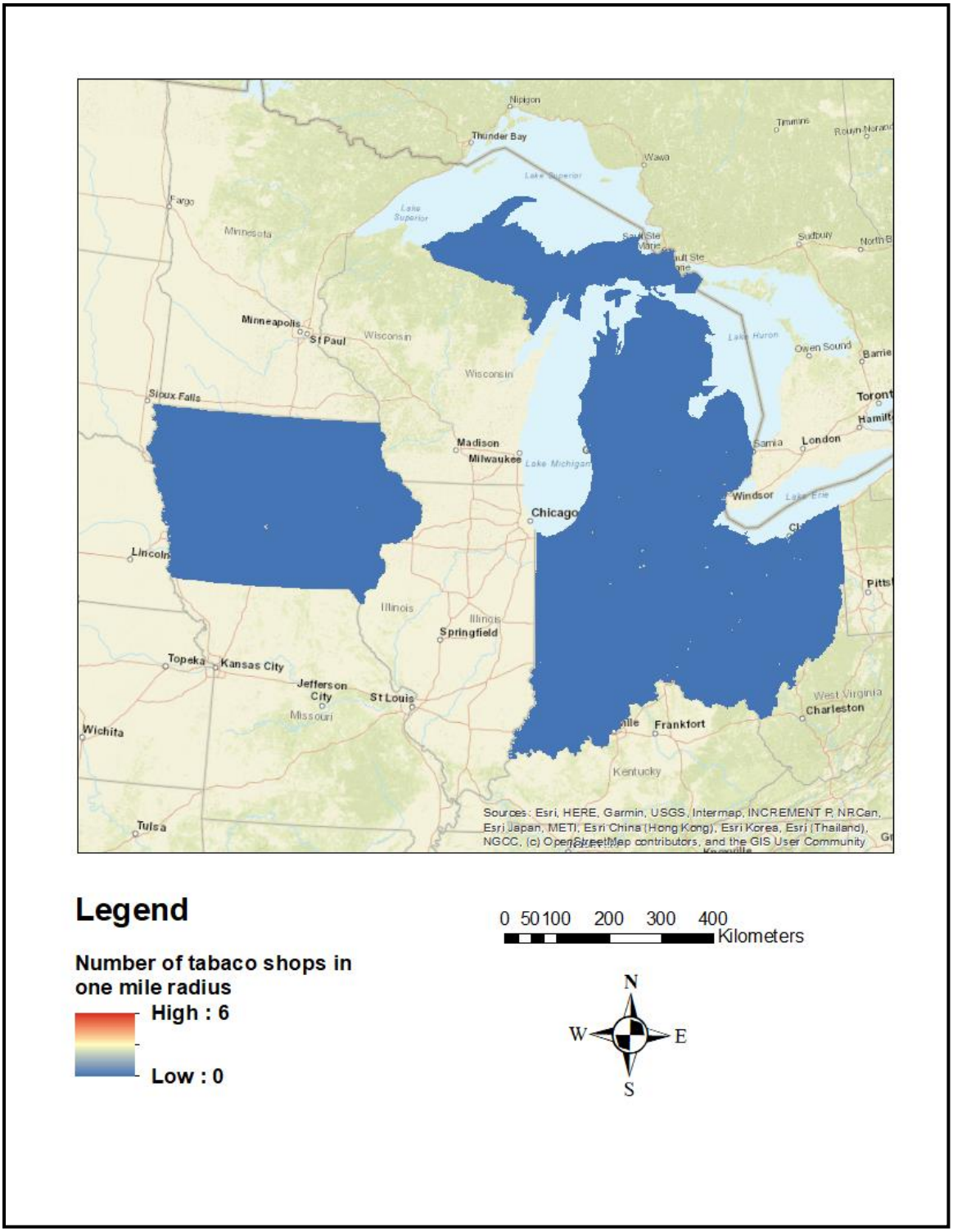


Figure 48. Model input: number of tobacco stores in one mile radius

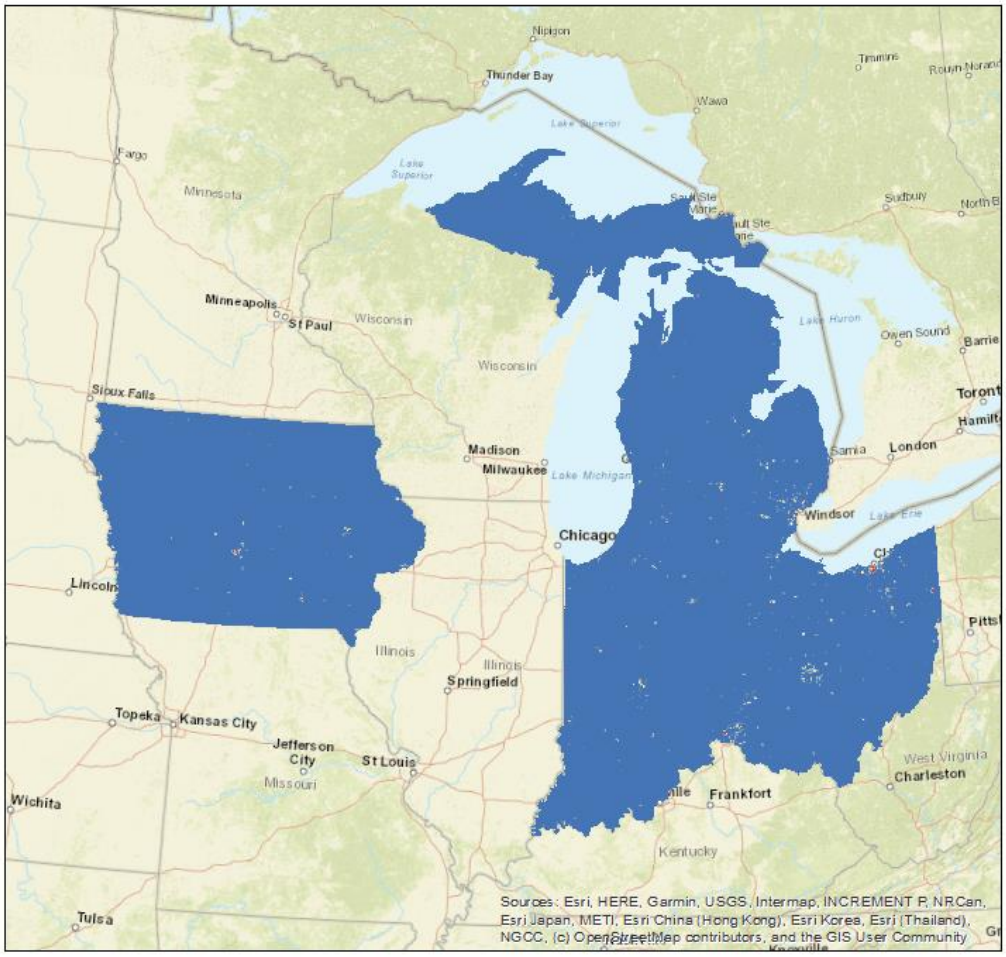


Figure 49. Number car shops in one mile radius

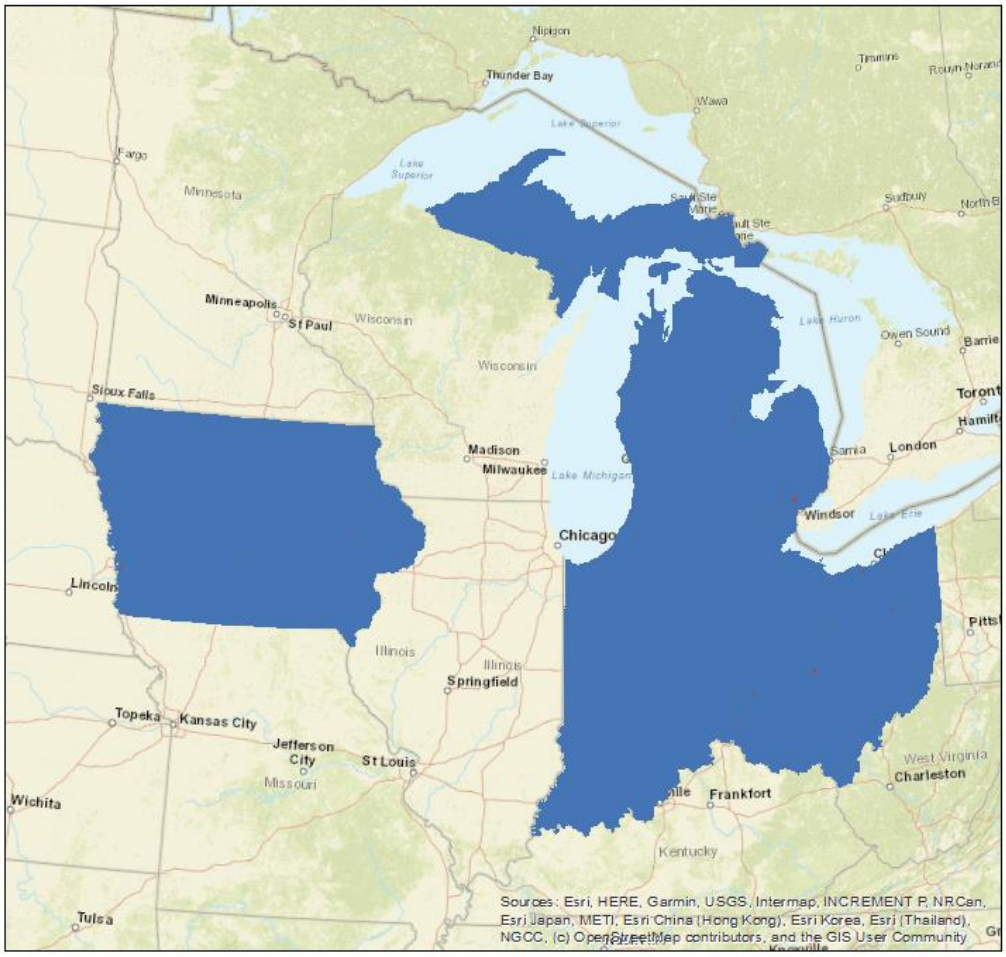


Figure 50. Model input: number children shops in one mile radius

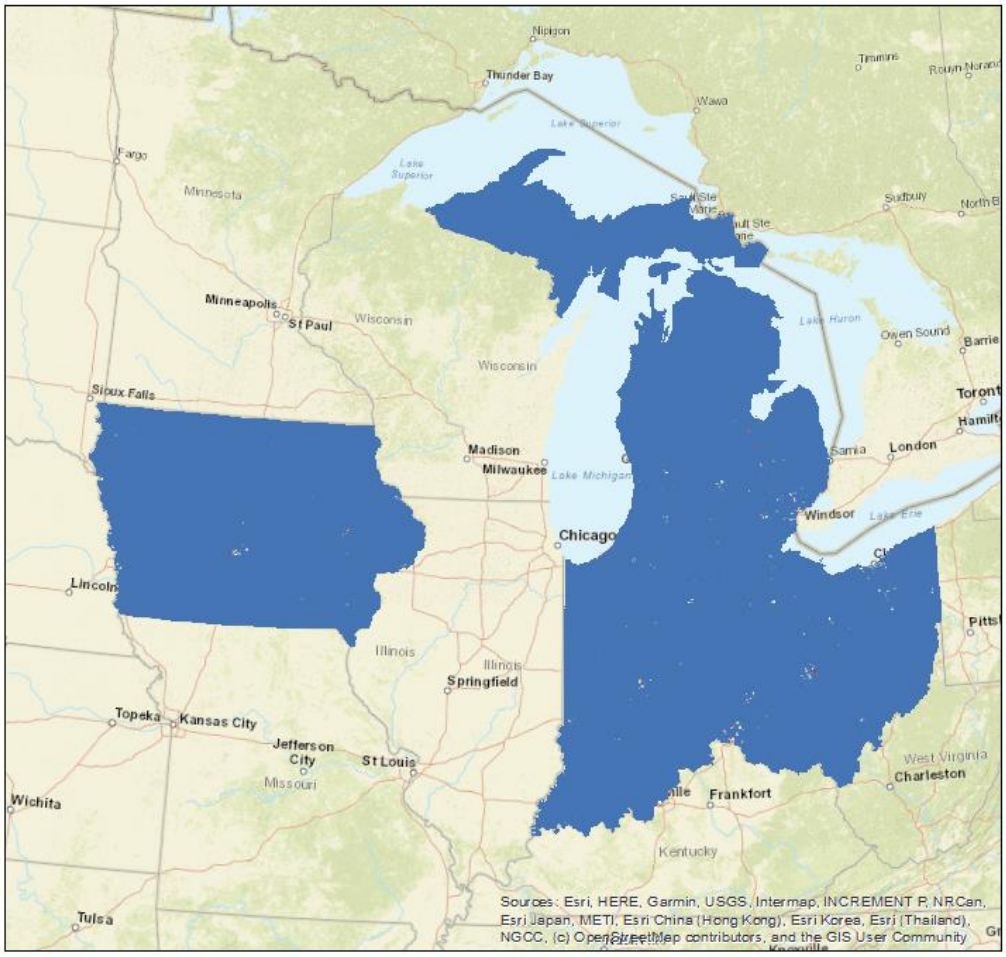


Figure 51. Model input: number clothes shops in one mile radius

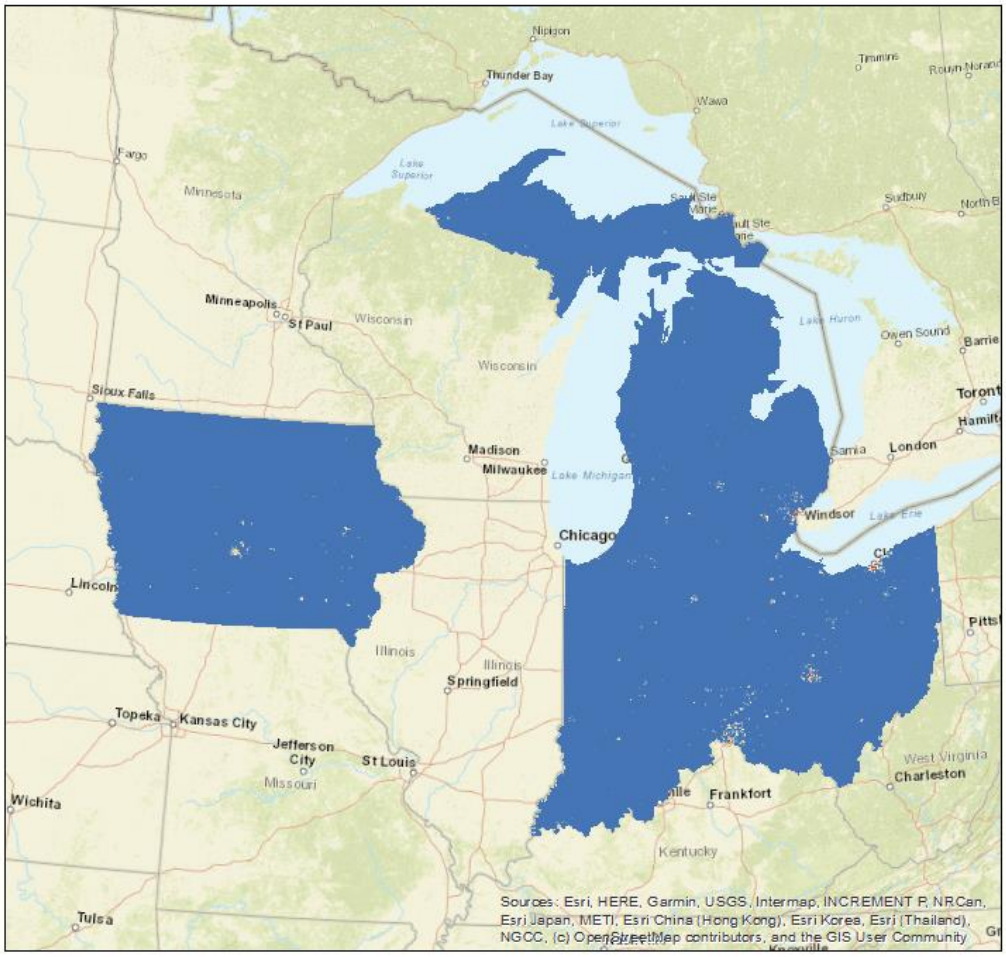


Figure 52. Model input: number convenience stores in one mile radius

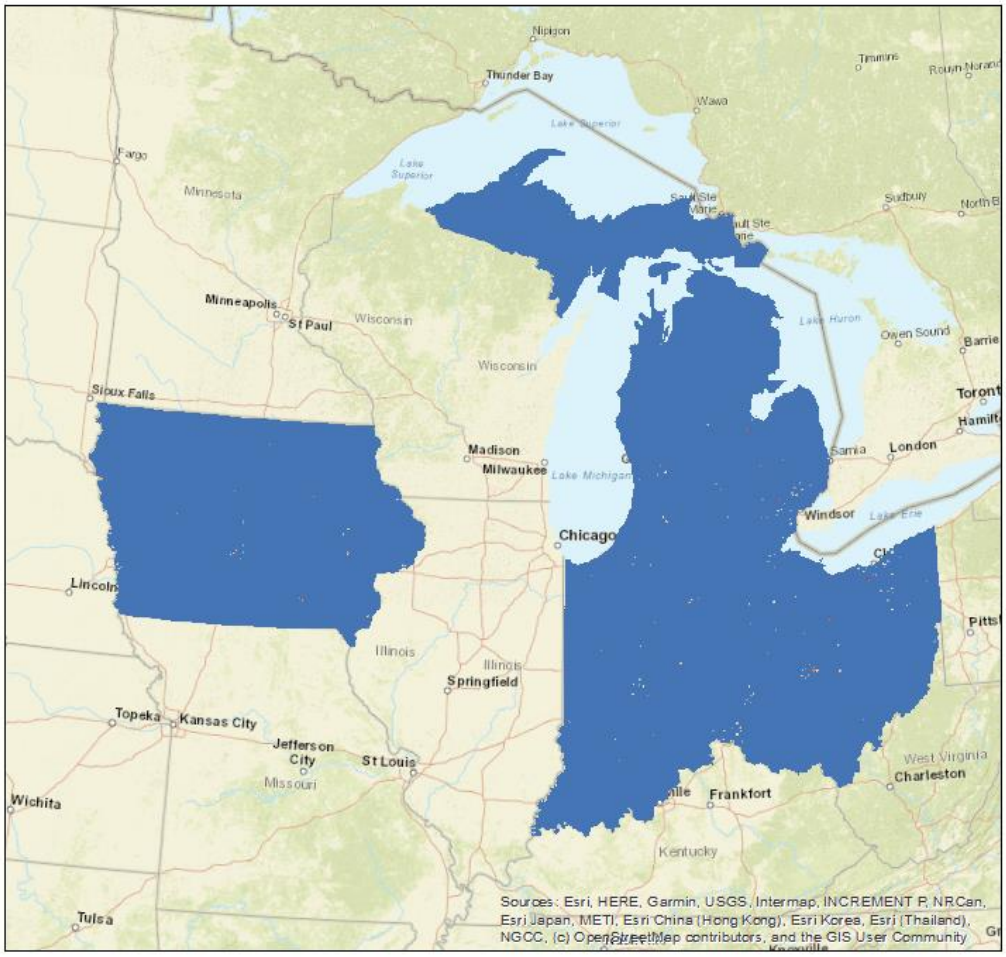


Figure 53. Model input: number department stores in one mile radius

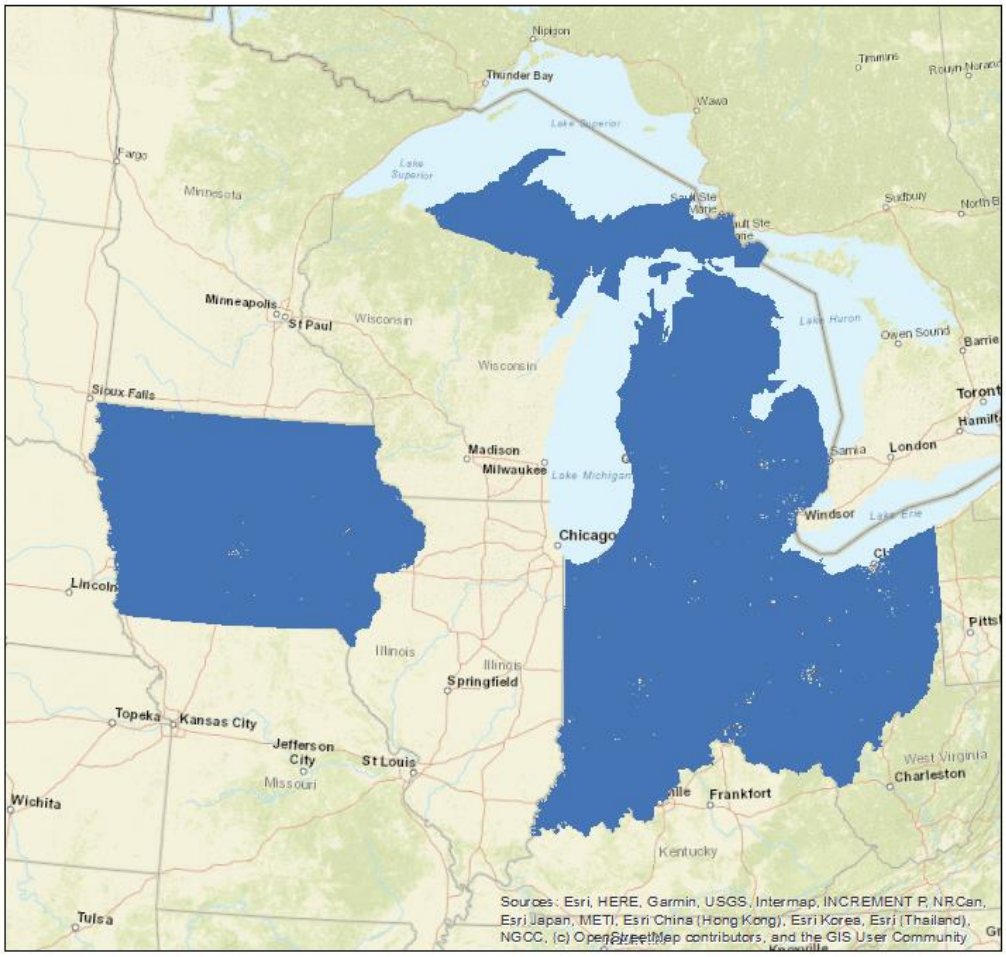


Figure 54. Model input: number electronics stores in one mile radius

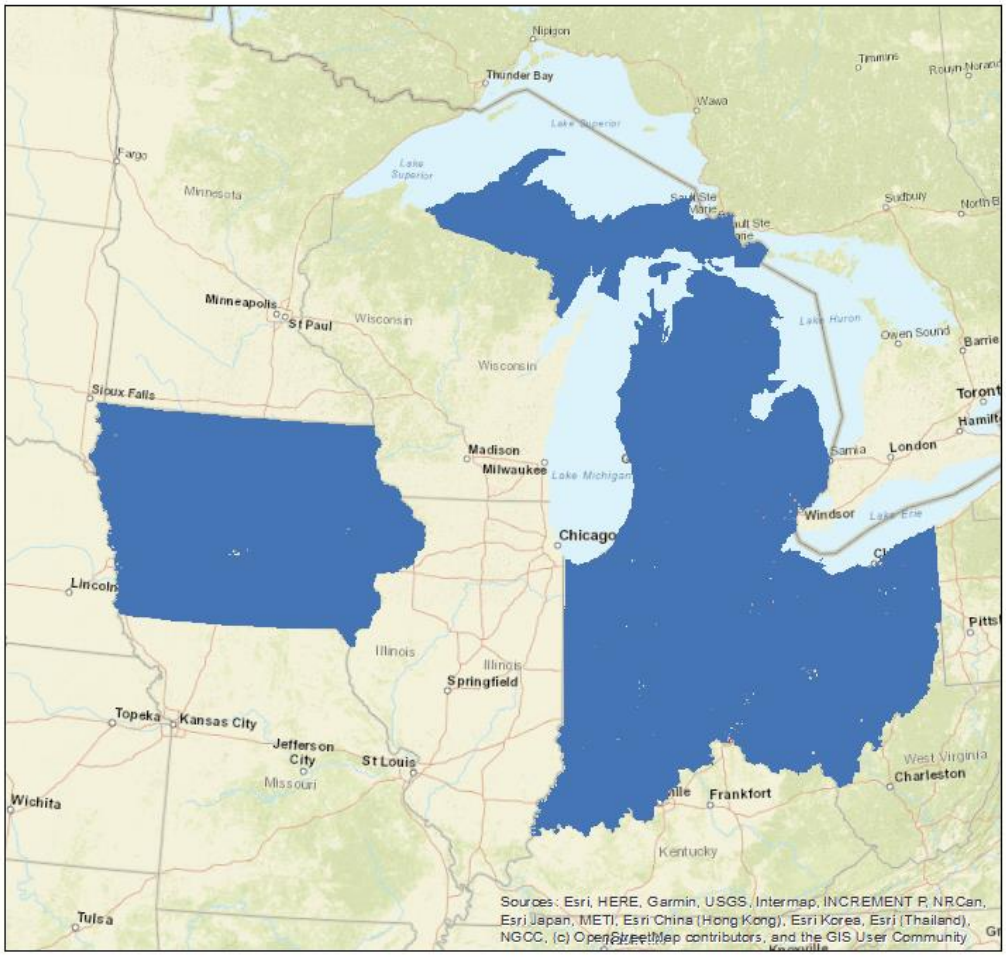


Figure 55. Model input: number entertainment stores in one mile radius

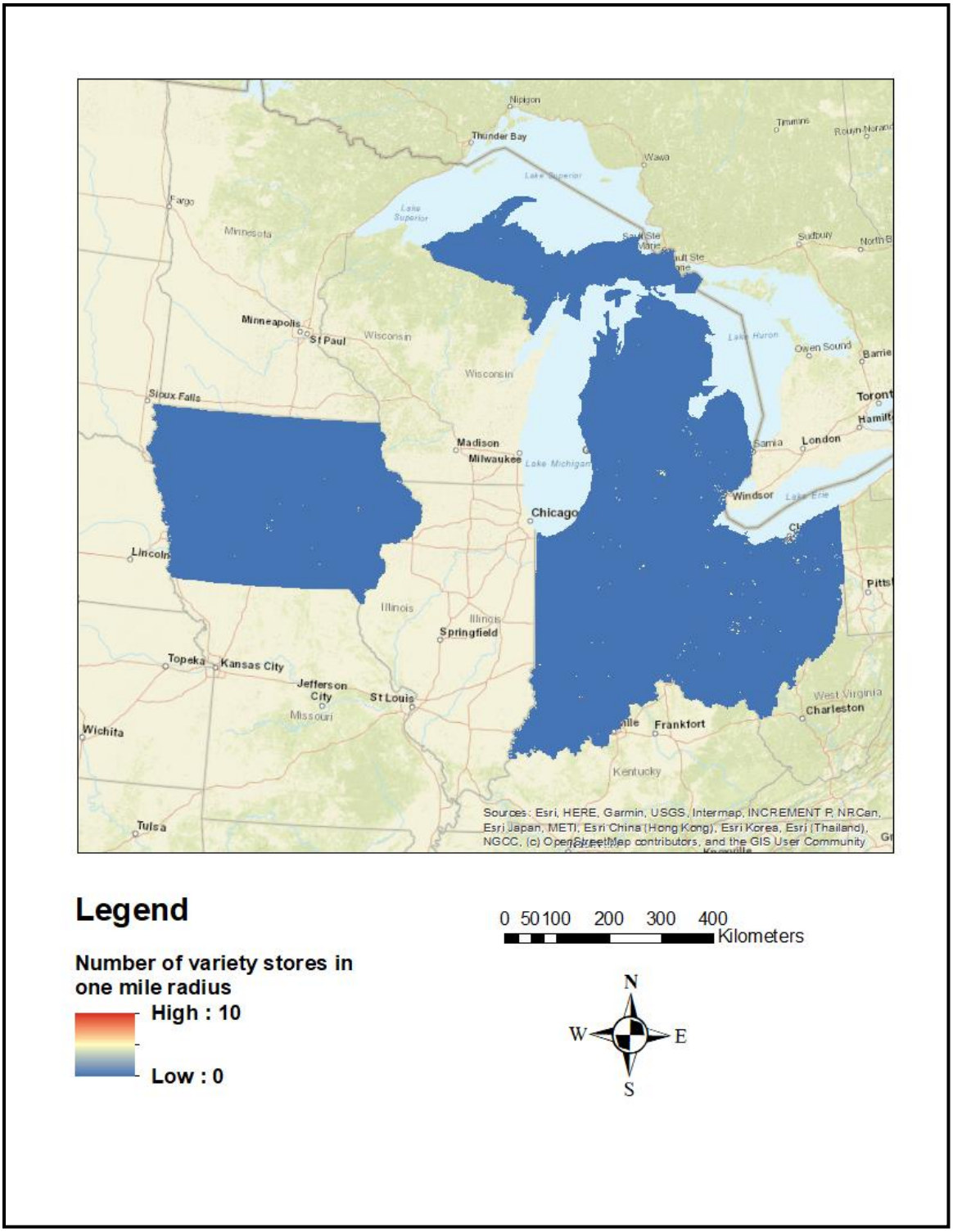


Figure 56. Model input: number of variety stores in one mile radius

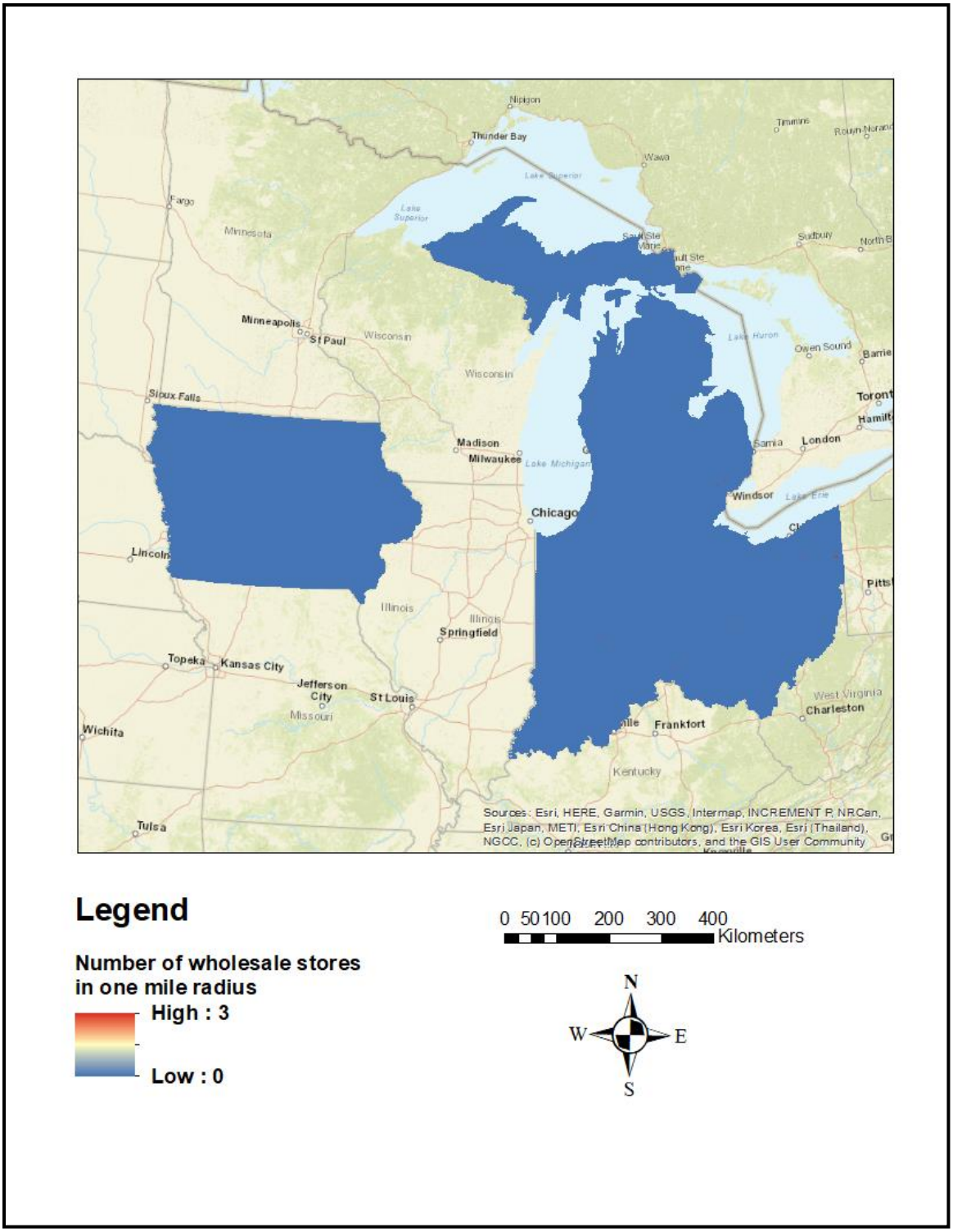


Figure 57. Model input: number of wholesale stores in one mile radius

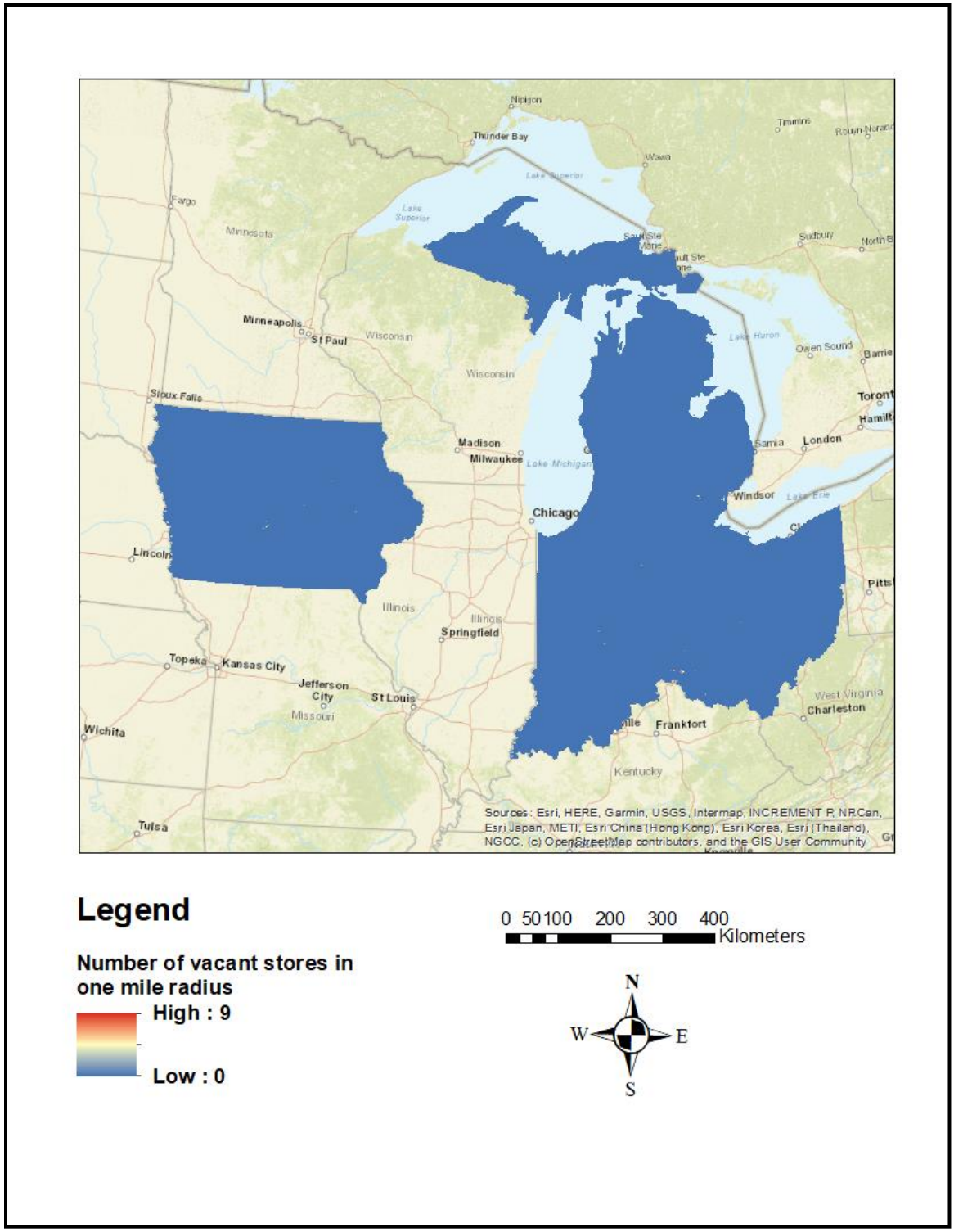


Figure 58. Model input: number of vacant stores in one mile radius

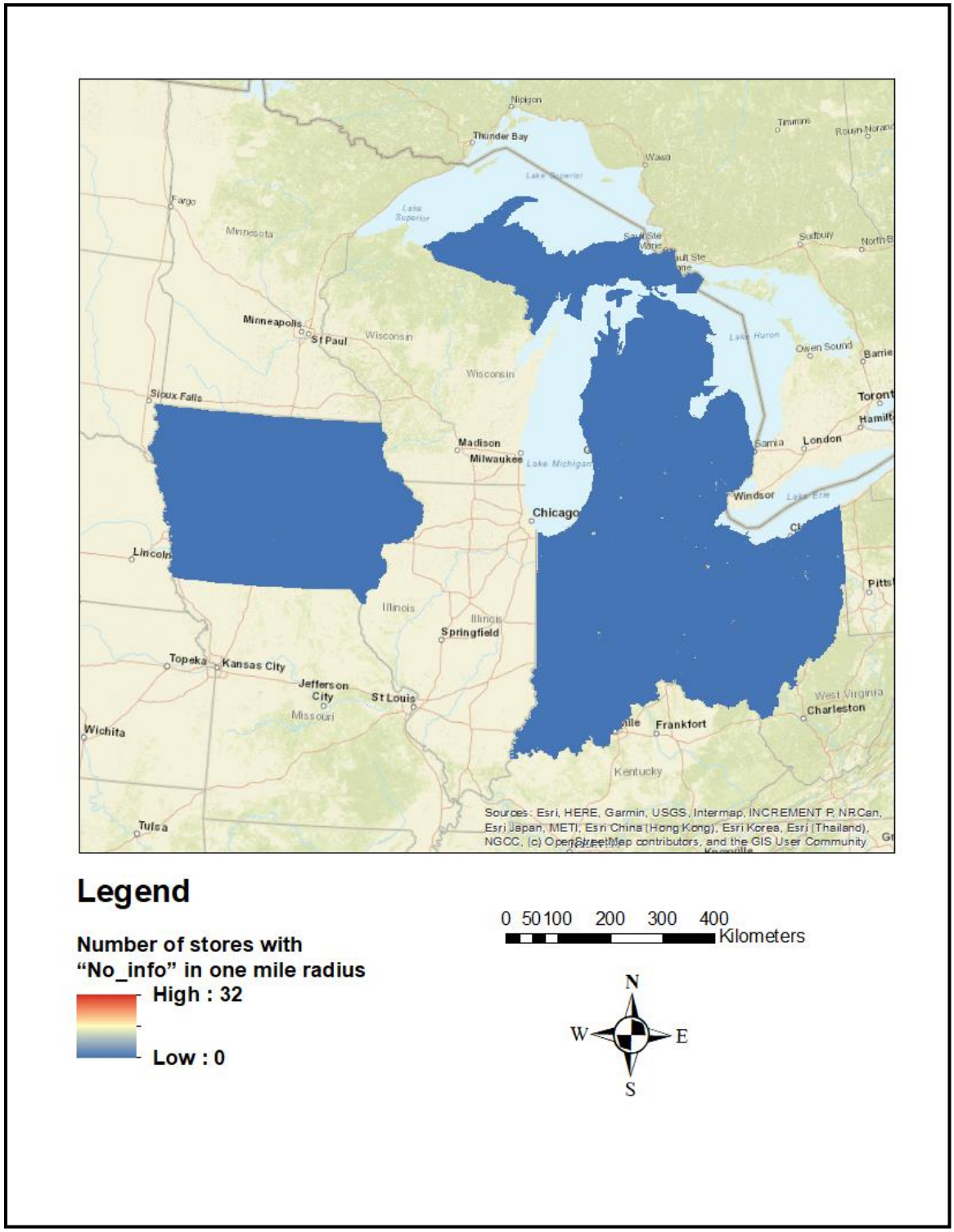


Figure 59. Model input: number of stores with "No info" status in one mile radius

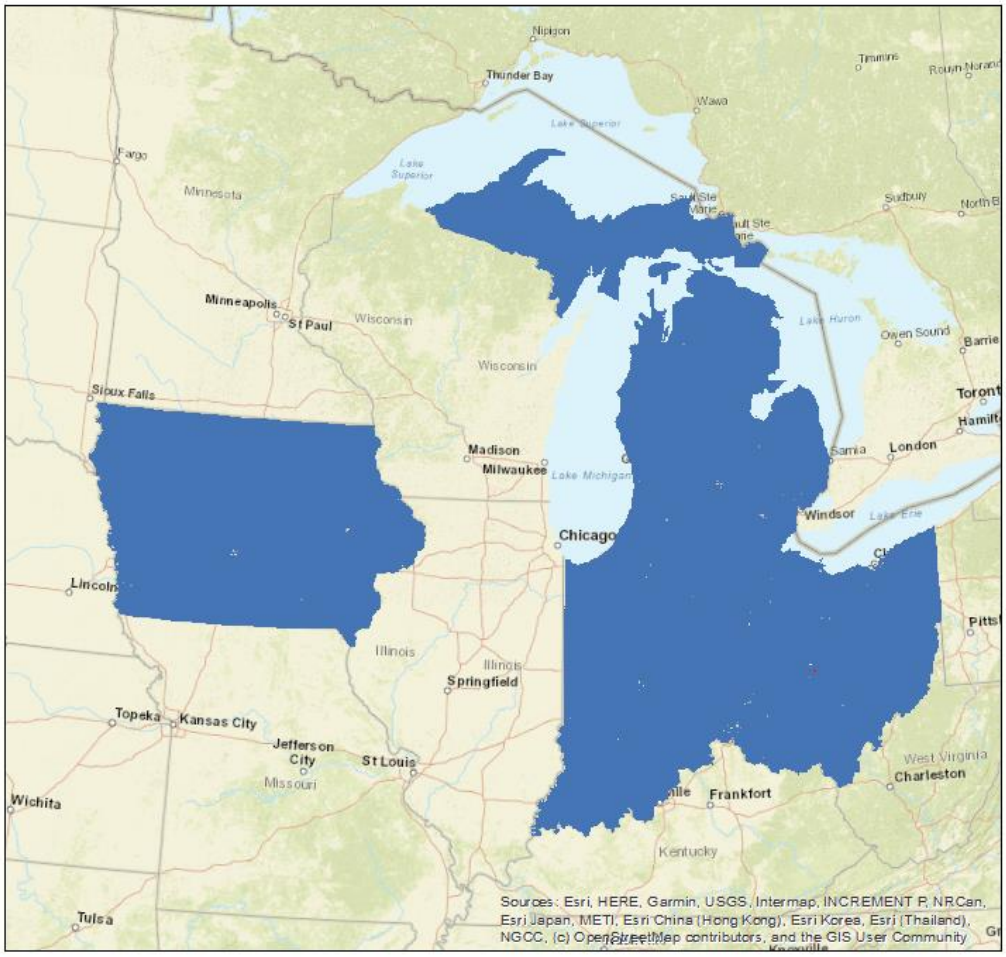


Figure 60. Model input: number of stores with status "Other" in one mile radius

Scripts

Script 1. Spearmen coefficient

To calculate the correlation, the following code was used:

```
>>from dask.distributed import Client
>>import joblib
>>import seaborn as sn
>>import pingouin as pg
>>with joblib.parallel_backend('dask'):
>>    df = pd.read_csv (r"D:\Thesis_low_resolution\df2.csv")
>>    # Calculate the pairwise Spearman correlation
>>     corr = pg.pairwise_corr(df, columns=['cafes',
>>     'fast_food',
>>     'Med_disp_inc',
>>     'Med_hh_inc',
>>     'Percent_of_AmInd_pop',
>>     'Percent_of_asian_pop',
>>     'Percent_of_black_pop',
>>     'Percent_of_hispanic_pop',
>>     'Percent_of_white_pop',
>>     'Percent_of_work_pop',
>>     'Population_Density',
>>     'Total_Population',
>>     'Parking_mask',
>>     'Restaurants_mask',
>>     'car',
>>     'services',
>>     'Children',
>>     'Clothes',
>>     'Convinience',
>>     'Department_store',
>>     'Electronics',
>>     'Entertaiment',
>>     'Furniture',
>>     'Gas',
>>     'Grocery',
>>     'Health',
>>     'Home',
>>     'Mall',
>>     'No_info',
```

```
>>     'Other',
>>     'Pet',
>>     'Sports',
>>     'Supermarkets',
>>     'Tabacoo',
>>     'Vacant',
>>     'Variety_store',
>>     'Wholesale'], method='spearman')
>>    # Sort the correlation by p-values and display the first rows
>>      #corr.sort_values(by=['p-unc'])[['X', 'Y', 'n', 'r', 'p-
unc']].head()
>>    df.corr().round(2)
>>    corrs = df.corr()
>>    mask = np.zeros_like(corrs)
>>    mask[np.triu_indices_from(mask)] = True
>>    sn.heatmap(corrs, cmap='Spectral_r', mask=mask, square=True,
vmin=-.4, >>vmax=.99)
>>    plt.title('Correlation matrix')
```

Script 2. VIF coefficient

```
>>vif_func<-function(in_frame,thresh=10,trace=T,...){
>>require(fmsb)
>>if(class(in_frame)       !=       'data.frame')       in_frame<-
data.frame(in_frame)
    #get initial vif value for all comparisons of variables
>>vif_init<-NULL
>>var_names <- names(in_frame)
>>for(val in var_names){
>>regressors <- var_names[-which(var_names == val)]
>>form <- paste(regressors, collapse = '+')
>>form_in <- formula(paste(val, '~', form))
>>vif_init<-rbind(vif_init,   c(val,   VIF(lm(form_in,   data   =
in_frame, ...))))
>>}
>>vif_max<-max(as.numeric(vif_init[,2]), na.rm = TRUE)
>>if(vif_max < thresh){
>>if(trace==T){ #print output of each iteration
>>prmatrix(vif_init,collab=c('var','vif'),rowlab=rep('',nrow(vif_
init)),quote=F)
>>cat('\n')
>>    cat(paste('All variables have VIF < ', thresh,', max VIF
',round(vif_max,2),
>>sep=''),'\n\n')
>>   }
>>return(var_names)
>>  }
>>else{
>>   in_dat<-in_frame
    #backwards selection of explanatory variables, stops when all
VIF values are below 'thresh'
>>while(vif_max >= thresh){
>>   vif_vals<-NULL
>>   var_names <- names(in_dat)
>>   for(val in var_names){
>>     regressors <- var_names[-which(var_names == val)]
>>     form <- paste(regressors, collapse = '+')
>>     form_in <- formula(paste(val, '~', form))
>>     vif_add<-VIF(lm(form_in, data = in_dat, ...))
>>     vif_vals<-rbind(vif_vals,c(val,vif_add))
>>   }
```

```
>>   max_row<-which(vif_vals[,2] == max(as.numeric(vif_vals[,2]),
na.rm = TRUE))[1]
>>   vif_max<-as.numeric(vif_vals[max_row,2])
>>   if(vif_max<thresh) break
>>   if(trace==T){ #print output of each iteration
>>
prmatrix(vif_vals,collab=c('var','vif'),rowlab=rep('',nrow(vif_va
ls)),quote=F)
>>    cat('\n')
>>    cat('removed: ',vif_vals[max_row,1],vif_max,'\n\n')
>>     flush.console()
>>   }
>>   in_dat<-in_dat[,!names(in_dat) %in% vif_vals[max_row,1]]
>>  }
>>   return(names(in_dat))
>> }
>>}
>>col<- vif_func(in_frame=colData,thresh=5,trace=T)
```