Spring 5-1-2020

Exploring Spatially Varying Relationships between Forest Fire and Environmental Factors in Fujian, China

QIANQIAN CAO
SUNY College of Environmental Science and Forestry, qcao03@syr.edu

EXPLORING SPATIALLY VARYING RELATIONSHIPS BETWEEN FOREST FIRE AND ENVIRONMENTAL FACTORS IN FUJIAN, CHINA

by

Qianqian Cao

A dissertation
submitted in partial fulfillment
of the requirements for the
Doctor of Philosophy Degree
State University of New York
College of Environmental Science and Forestry
Syracuse, New York
May 2020

Department of Sustainable Resources Management

Approved by:
Lianjun Zhang, Major Professor
Gary Scott, Chair, Examining Committee
Christopher Nowak, Department Chair
S. Scott Shannon, Dean, The Graduate School

Acknowledgements

I would like to thank my advisor, Dr. Lianjun Zhang, for his support, patience, and encouragement throughout my graduate studies. When I was a master student, he introduced me into the interesting field of statistics. During my Ph.D. study, he guided me to think like a scientist, solve problems in a scientific process, and provide insights on the workings of academic research. Whenever I feel lost and/or frustrated in research, his advice was essential to the completion of this dissertation. I appreciate his valuable advice and constant encouragement during my studies. I really enjoyed every moment of the meetings and discussions with him. Dr. Zhang's guidance and mentorship not only improved my research ability, but also influenced other aspects of my life.

Thanks to my thesis committee members, Dr. Stephen Stehman, Dr. Lindi Quackenbush, Dr. Diane Kiernan, Dr. Eddie Bevilacqua, and Dr. Gary Scott (Chair) for their encouragement and constructive inputs. Thanks to Dr. Stephen Stehman for the valuable discussions on variable selections. Thank you again to Dr. Lindi Quackenbush for providing me the seminars and literatures related to my research. Thanks to Dr. Diane Kiernan for encouraging me to be a good teaching assistant, an instructor, and a mom. Thanks to Dr. Eddie Bevilacqua for serving as my examiner in the midst of pressing affairs. Thanks to Dr. Gary Scott for holding my defense as the Chair. I also thank Dr. Futao Guo, who provided the data of forest fire in Fujian province, China for this dissertation. In addition, I am very thankful to my friends at ESF and Syracuse University, Siqi Li, Ying Zhou, Chuan He, and Ming Li, for their kindness, friendship, and support.

The Year 2020 is unusual, stressful, and challenging. Our lives, how we learn, how we work, how we consume, and how we socialize and communicate, have to change due to the spread of COVID-19 pandemic. The courses in colleges are switched to online teaching and the defense of my dissertation will be conducted online too. While we are separated by distance, I do miss each and every one of you at ESF. Further, I am most grateful to my parents for helping me with two babies, Clark (two-years old) and Lewis (four-month old), giving me considerable support and spiritual comfort throughout my Ph.D. studies. My dear husband, Fengbin, your love, support, understanding, and constant patience during my pursuit of Ph.D. degree make the completion of this dissertation possible. Words would never say how grateful I am to all my family. I am blessed to have such a lovely and caring family, standing beside me with their love and unconditional support.

Table of Contents

List of Tables

List of Figures

Abstract

Q. Cao. Exploring Spatially Varying Relationships between Forest Fire and Environmental Factors in Fujian, China, 145 pages, 13 tables, 22 figures, 2020. APA style guide used.

In recent decades, the occurrence of forest fires has risen in the world and led to significant, long-lasting impacts on ecological, social, and economic systems. Along with the traditional tools for fire prediction, statistical modeling has been playing an important role in understanding the nature of forest fires and providing guidelines for decision making of fire prevention and management. In this dissertation, a large data set was collected from 2001 to 2016 in Fujian province, China, including the occurrence of forest fires and many environmental factors. I developed spatial generalized linear models and spatial quantile models under the framework of geographically weighted regression (GWR) to investigate the relationships between the counts and proportion or rate of forest fires and driving topographical, meteorological, human, vegetation, and land coverage factors. The corresponding global models were used as the benchmarks for model comparisons. These spatial models included: (1) geographically weighted Poisson and geographically weighted negative binomial models designed for the counts of forest fires; (2) geographically weighted quantile models for the counts of forest fires at different quantiles or risk levels; and (3) geographically weighted beta model for the proportion or rate of forest fires. The results indicated that the observed forest fires were highly likely to occur in lower elevation, smaller aspect index (meaning stronger sunlight), heavier precipitation, smaller population density, less vegetation, wider grassland, and/or less cropland, while other environmental factors varied greatly with the forest fire occurrence. This study showed the great superiority of these GWR models to the corresponding global models in terms of characterizing the spatial nonstationary relationships, producing better model fitting and prediction, providing a more complete view on the spatial distribution of forest fires, and highlighting the risky local "hot spots" of forest fires as well as environmental factors across the Fujian province, China. Hopefully, the more detailed and localized information would help and assist the forest and fire managers to better understand the behavior of forest fires and influences of the environmental factors across the study area. Thus, the government agencies can make wiser and better decisions on where and what the fire management and prevention should be focused on with reduced economic expenses and improved the efficiency of forest fire management.

Key Words: occurrence of forest fire; driving factors; geographically weighted Poisson regression; geographically weighted negative binomial regression; geographically weighted quantile regression, geographically weighted beta regression.

Q. Cao
Candidate for the degree of Doctor of Philosophy, May 2020
Lianjun Zhang, Ph.D.
Department of Sustainable Resources Management
State University of New York College of Environmental Science and Forestry,
Syracuse, New York

Chapter I: Introduction

Overview

Fire in ancient Greeks was considered as one of the classical elemental forces along with water, earth, and air (Thomas *et al*. 2010). Fire ignited in forests spread out over a large area of natural lands. On one hand, forest fires may benefit the ecosystems (Thomas *et al*. 2010) by cleaning fallen tree limbs and brushes, releasing the nutrition back to environment, and stimulating new generation of plant life. This cycle is repeated almost endless and slowly changing the ecosystem. On the other hand, forest fires usually interrupt wildlife habitats and landscape, release carbon dioxide into the atmosphere, destroy homes and communities, and pollute air with emissions harmful to human health (Scott *et al.* 2013). Forest fire is an essential element useful for ecosystems, but it can threaten lives, structures, infrastructure, dependent economics, and our way of life if it is unmanaged or uncontrolled (Thomas *et al*. 2010).

In order to help suppress wildland fires, reduce fire risks, and minimize damages to ecosystem and human beings, statistical modeling has been playing an important role in understanding the nature of forest fires and providing guidelines for decision making of wildfire prevention and management. Pastor *et al.* (2003) reviewed the most important works on the mathematical models of wildfires since 1940s. According to their nature, these models can be classified into theoretical, empirical, and semi-empirical models. The theoretical models are generated from the laws controlling fluid mechanics, combustion, and heat transfer. The empirical models are composed of statistical correlations between wildfire and impact factors from experiments or historical wildfire studies. The semi-empirical models are proposed from general theoretical expression and validated through experimentation. Within the three categories, the empirical models are the focus of this dissertation. The details of other two types of forest fire models can be found in the literature (Pastor *et al.* 2003).

Predicting the probability or count of forest fire occurrence are two major aspects in empirical models. As the earliest fire index empirical model developed in 1966, the McArthur's fire-danger meter has been operating in Australia. Its equations were re-fitted later by Noble *et al.* (1980) applying the linear and exponential functions of relative humidity, air temperature, average wind velocity, and drought factor to calculate the fire danger index. Additionally, Forestry Canada Fire Danger Group (1992) built the Canada forest fire behavior prediction system based on the linear correlation relationships between forest fire behavior and weather, fuel, topography, and elapsed time (De Groot 1993).

Development of computing power in recent decades has led to wide applications of regression methods to investigate the relationships between forest fire and driving factors. To predict the number of wildfires, logistic regression and multiple linear regression were widely applied (Vega Garcia *et al.* 1995; Andrews and Bradshaw 1997; Brillinger *et al.* 2006; Guo *et al.* 2016a; Syphard *et al.* 2007; Sebastián-López *et al.* 2008; Oliveira *et al.* 2012). However, both logistic and multiple linear regression models have to transfer the forest fire count from discrete to continuous. The scale change of the response variable is difficult in model interpretation. Instead of transforming the discrete count data to a continuous scale, the expected number of forest fires can be estimated via generalized linear models. Poisson regression has been considered a reliable method for forest fire counts since the 1970s (Cunningham and Martell 1973; Dayananda 1977; Gill *et al*. 1987; Mandallaz and Ye 1997; Wotton *et al.* 2003, 2010). Additionally, negative binomial model has also been applied in forest fire modeling when the forest fire data (dependent variable) are dispersed (Cardille *et al.* 2001; Xiao *et al*. 2011; Chas-Amil *et al*. 2015).

Predicting the occurrence probability of forest fires is another essential key in statistical modeling. Logistic regression is the most popular technique used to quantify the relationships and predictions between the occurrence probability of forest fires and potential explanatory variables (Preisler *et al*. 2004; Preisler and Westerling 2007; Lozano *et al*. 2007; Hoyo *et al.* 2011; Chang *et al.* 2013; Guo *et al.* 2016a, 2016b). However, it is only applicable when the response variable is binary, i.e., in the outcome there are two possibilities, assigned the value 0 or 1. Therefore, the logistic regression cannot be used when the response variable, the probability of fire occurrence, is continuous within the interval (0, 1). In this case, beta regression is an alternative technique to model probability, proportion, or rate, introduced by Ferrari and Cribari-Neto (2004). These generalized liner models, including logistic, Poisson, negative binomial, and beta regression, target on the “mean or average” relationships between response variable (e.g., forest fire) and predictor variables (e.g., environmental factors). Therefore, these models provided the explanation and prediction on the average level of the forest fires, given the values of the predictors.

When forest fire managers have great interests in the impacts of driving factors at various risk levels beyond an average fire risk, quantile regression is a good choice for modeling the whole range of forest fire risks. Introduced by Koenker and Bassett (1978), quantile regression is viewed as an extension of classical estimation of conditional mean models to the estimation of an ensemble of models for several conditional quantile functions (Koenker and Hallock 2001). It computes several different regression curves corresponding to the various percentage points of the distribution (Mosteller and Tukey 1977). Applying the quantile regression to forest fire would provide a more complete picture of driving factors.

Because spatial dependences and heterogeneity exist across forest ecosystems, many statistical models have been extended to the framework of geographically weighted regression (GWR) to investigate the spatially varying relationships between response variables and predictors in various study fields over the last 20 years, such as remote sensing (Foody *et al.* 2003), landscape pattern (Su *et al.* 2012), deforestation (Jaimes *et al*. 2010), urban ecology (Su *et al.* 2014), AND forestry (Zhang and Shi 2004; Wang *et al*. 2005; Shi *et al.* 2006; Guo *et al.* 2008; Kimsey *et al.* 2008; Salvati *et al.* 2015; Segovia *et al.* 2016). The major advantages of the spatially localized regression models are: (1) develop a model for each geographic location in the study area, focusing on local exceptions, statistics, and relationships between response variable and predictor variable rather than global regularities; (2) produce better model fitting and location-specific model predictions for the response variable of interest; (3) identify "hot spots" where local clustering and nonstationarity of the response variables around a particular location; and (4) provide mappable model statistics and parameters that can be used to visualize the spatial patterns of the "local" relationships across the study area (Fotheringham *et al*. 2002; Wheeler and Páez 2010; Chen and Yang 2012). The forest coverage in Fujian province, China ranks the highest with high annual forest fire incidences (Guo *et al*. 2017). However, the statistical analysis of influencing factors on forest fires has been mainly focused on the boreal forests in northern China (Liu *et al.* 2012; Wu *et al.* 2014; Guo *et al.* 2015; Guo *et al.* 2016b; Guo *et al.* 2016c), which results in a less informative fire management plan specifically designed for subtropical regions like the Fujian province. Further, there are limited applications of the spatially localized regression techniques such as GWR to model forest fires (probability and/or count) and topographical, meteorological, human factors, and vegetation and land coverage across the study areas (Guo *et al*. 2016a; Guo *et al.* 2016b; Guo *et al*. 2017).

The main objective of this dissertation was to develop statistical models for exploring the spatially varying relationships between the occurrences of forest fires and driving factors in the sub-tropic region of China. Specifically, the statistical models included: (1) geographically weighted Poisson and geographically weighted negative binomial models designed for the count of forest fires; (2) geographically weighted quantile models for the count of forest fires at different quantiles or risk levels; and (3) geographically weighted beta model for the proportion or rate of forest fires. These spatially localized models were aimed at finding the spatial patterns of influential environmental factors on the occurrence of forest fires in the study area.

It is known that the smoothing process of estimating model coefficients for a global regression can overlook the interesting geographical features in the relationships between response variable and predictor variables (Su *et al.* 2017). Therefore, it is expected that the spatially localized generalized linear regression models and quantile regression models under the framework of GWR would show a great superiority to the corresponding global models in terms of characterizing the spatial nonstationary relationships, producing better model fitting and predicting performances, and providing a more critical and adequate look on the influential factors of forest fire at existing ‘whole’ maple statistics. We expected the newly developed GWR models improved the analyses of risky factors of forest fires, and hopefully, would provide wiser and better insight into the forest fire mapping, understanding, prevention, and management based on the local character prospects.

References

Andrews P.L., Bradshaw L.S., 1997. Fires: Fire information retrieval and evaluation system - a program for fire danger rating analysis. USDA Forest Service, Intermountain Research Station General Technical Report INT-367. 64 pp. (Ogden, UT).

Brillinger D.R., Preisler H.K., Benoit J.W., 2006. Probabilistic risk assessment for wildfires. *Environmetrics* 17(6): 623-633.

Cardille J.A., Venture S.J., Turner M.G., 2001. Environmental and social factors influencing wildfires in the upper Midwest, United States. *Ecological Applications* 11(1): 111-127.

Chang, Y., Zhu, Z., Bu, R. Chen H., Feng Y., Li Y., Hu Y., Wang Z., 2013. Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landscape Ecology* 28: 1989-2004.

Chas-Amil M.L., Prestemon J.P., McClean C.J., Touza J., 2015. Human-ignited wildfire patterns and responses to policy shifts. *Applied Geography* 56: 164-176.

Chen Y., Yang T., 2012. SAS macro programs for geographically weighted generalized linear modeling with spatial point data: Applications to health research. *Computer Methods and Programs in Biomedicine* 107(2): 262-273.

Cunningham A.A., Martell D.L., 1973. A stochastic model for the occurrence of man-caused forest fires. *Canadian Journal of Forest Research* 3: 282-287.

Dayananada P.W.A., 1977. Stochastic models for forest fires. *Ecological Modelling* 3: 309-313.

De Groot W.J., 1993. Examples of fuel types in the Canadian Forest Fire Behavior Prediction (FBP) System. Poster with text, Forestry Canada, Northern Forestry Centre, Edmonton, Alberta.

Ferrari, S.L.P, Cribari-Neto, F., 2004. Beta Regression for modelling rates and proportions. *Journal of Applied Statistics* 31(7): 799-815.

Foody G. M., Boyd D. S., Cutler, M. E. J., 2003. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sensing of Environment* 85(4): 463-474.

Forestry Canada Fire Danger Group,1992. Development and structure of the Canadian forest fire behavior prediction system. Canadian Department of Forestry, Inference Reprehensive. ST-X-3.

Fotheringham A.S., Brunsdon C., Charlton M., 2002. Geographically weighted regression. Wiley, New York. 284 p.

Gill A.M., Christian K.R., Moore P.H.R., 1987. Bush fire incidence, fire hazard and fuel reduction burning. *Australian Journal of Ecology* 12: 299-306.

Guo L., Ma Z., Zhang L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. *Canadian Journal of Forest Research* 38(9): 2526-2534.

Guo F., Wang G., Su Z., Liang H., Wang, W., Lin F., Liu A., 2016a. What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire* 25: 505-519.

Guo F., Selvalakshmi S., Lin F., Wang G., Wang W., Su Z., 2016b. Geospatial information on geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese boreal forest. *Canadian Journal of Forest Research* 46: 582-594.

Guo F., Zhang L., Jin S., Tigabu M., Su, Z., Wang W., 2016c. Modeling anthropogenic fire occurrence in the boreal forest of China using logistic regression and Random Forest. *Forests* 7: 250-264.

Guo, F., Su, Z., Wang, G., Sun, L., Tigabu, M., Yang, X., Hu, H., 2017. Understanding fire drivers and relative impacts in different Chinese forest ecosystems. *The Science of the Total Environment* 605: 411-425.

Liu Z., Yang J., Chang Y., Weisberg P. J., He H. S., 2012. Spatial patterns and drivers of fire occurrence and its future trend under climate change in a boreal forest of northeast China. *Global Change Biology* 18: 2041-2056.

Hoyo V.L., Martín Isabel M., Martínez Vega F., 2011. Logistic regression models for human-caused wildfire risk estimation: analyzing the effect of the spatial accuracy in fire occurrence data. *European Journal of Forest Research* 130: 983-996.

Jaimes N. B. P., Sendra J. B., Delgado M. G., Plata R. F., 2010. Exploring the driving forces behind deforestation in the state of Mexico (Mexico) using geographically weighted regression. *Applied Geography* 30: 576–591.

Kimsey M., Moore J., McDaniel P., 2008. A geographically weighted regression analysis of Doulgas-Fir site index in North Central Idaho. *Forest Science* 54(3): 356-366.

Koenker R., Bassett J.R.G., 1978. Regression quantiles. *Econometrical. Journal of the Econometric Society* 46: 33-50.

Koenker R., Hallock K., 2001. Quantile Regression. *The Journal of Economic Perspective* 15 (4): 143-156.

Lozano F.J., Suárez-Seoane S., de Luis E., 2007. Assessment of several spectral indices derived from multi-temporal landsat data for fire occurrence probability modelling. *Remote Sensing of Environment* 107: 533-544.

Mandallaz D., Ye R., 1997. Prediction of forest fires with Poisson models. *Canadian Journal of Forest Research* 27: 1685-1694.

McArthur A.G., 1966. Weather and grassland fire behavior. Forest Research Institute, Forest and Timber Bureau of Australia. Leaflet No. 100.

Mosteller F., Tukey J., 1977. Data Analysis and Regression: A Second Course in Statistics. Reading, Mass.: Addison-Wesley.

Noble I.R., Bary G.A.V., Gill A.M., 1980. McArthur's fire-danger meters expressed as equations. *Australian Journal of Ecology* 5: 201-203.

Oliveira S., Oehler F., San-Miguel-Ayanz J., Camia A., Pereira J., 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and Random Forest. *Forest Ecology and Management* 275: 117-129.

Pastor E., Zarate L., Planas E., Arnaldos J., 2003. Mathematical models and calculation systems for the study of wildland fire behavior. Progress in Energy and Combustion Science 29: 139-153.

Preisler H.K., Brillinger D.R., Burgan R.E., Benoit J.W., 2004. Probability based models for estimation of wildfire risk. *International Journal of Wildland Fire* 13(2): 133-142.

Preisler H.K., Westerling A.L., 2007. Statistical model for forecasting large wildfire events in western United States. *Journal of Applied Meteorology and Climatology* 46: 1020-1030.

Salvati L., Biasi R., Carlucci M., Ferrara A., 2015. Forest transition and urban growth: exploring latent dynamics (1936-2006) in Rome, Italy, using a geographically weighted regression and implications for coastal forest conservation. *Rendiconti Lincei* 26: 577-585.

Scott J.H., Thompson M.P., Calkin D.E., 2013. A wildfire risk assessment framework for land and resource management. General Technical Report RMRS-GTR-315. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, p 83.

Segovia M.Q., Ruiz S.C., Drápela K., 2016. Comparison of height-diameter models based on geographically weighted regressions and linear mixed modelling applied to large scale forest inventory. *Forest Systems* 25(3): e076, 11 pages.

Sebastián-López A., Salvador-Civil R., Gonzalo-Jiménez J., San-Miguel Ayanz J., 2008. Integration of socio-economic and environmental variables for modelling long-term fire danger in Southern Europe. *European Journal of Forest Research* 127: 149-163.

Shi H., Zhang L., Liu J., 2006. A new spatial-attribute weighted function for geographically weighted regression. *Canadian Journal of forest research* 36(4): 996-1005.

Su, S., Xiao, R., Zhang, Y., 2012. Multi-scale analysis of spatially varying relationships between agricultural landscape patterns and urbanization using geographically weighted regression. *Applied Geography* 32(2): 360-375.

Su, S., Li, D., Hu, Y., Xiao, R., Zhang, Y., 2014. Spatially non-stationary response of ecosystem service value changes to urbanization in shanghai, China. *Ecological Indicators* 45(5): 332-339.

Su S., Gong Y., Tan B., Pi J., Weng M., Cai Z., 2017. Area social deprivation and public health: analyzing the spatial non-stationary associations using geographically weighted regression. *Social Indicators Research* 133(3): 819-832.

Syphard, A.D., Radeloff, V.C., Keely, J.E., Hawbaker, R.J., Clayton, M.K., Stewart, S.I., Hammer, R.B., 2007. Human influence on California Fire Regimes. *Ecological Applications* 17: 1388-1402.

Thomas P.A., McAlpine R.S., Hobson P., 2010. Fire in the Forest. Cambridge University Press.

Vega Garcia C., Woodard P.M., Titus S.J., Adamowicz W.L., Lee B.S., 1995. A logit model for predicting the daily occurrence of human caused forest fires. *International Journal of Wildland Fire* 5(2): 101-111.

Wang Q., Ni J., Tenhunen J., 2005. Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. *Global Ecology and Biography* 14(4): 379-393.

Wheeler D., Pa´ez A., 2010. Geographically weighted regression. In M.M. Fisher and A. Getis (Eds.), *Handbook of applied spatial analysis: Software tools, methods and applications* (pp. 461-486). Heidelberg: Springer.

Wotton B.M., Martell, D.L., Logan, K.A., 2003. Climate change and people-caused forest fire occurrence in Ontario. *Climatic Change* 60(3): 275-295.

Wotton B.M., Nock C.A., Flannigan M.D., 2010. Forest fire occurrence and climate change in Canada. *International Journal of Wildland Fire* 19(3): 253-271.

Wu Z., He H.S., Yang J., Liu Z., Liang, Y., 2014. Relative effects of climatic and local factors on fire occurrence in boreal forest landscapes of northeastern China. *Science of the Total Environment* 493: 472-480.

Xiao Y., Ju H., Zhang X., Ji P., 2011. Relationship between fire-danger weather and forest fire in Qiannan area. *Scientia Silvae Sinicae* 47: 128-133. (in Chinese)

Zhang L. Shi H., 2004. Local modeling of tree growth by geographically weighted regression. *Forest Science* 50(2): 225-244.

Chapter II: Forest Fire Occurrence Modeling and Analysis of Driving Factors Using Global and Geographically Weighted Regression

Abstract. We applied global Poisson, global negative binomial (NB), geographically weighted Poisson regression (GWPR), and geographically weighted negative binomial regression (GWNBR) to investigate the spatially varying relationships between forest fire count and topographical, meteorological, human, vegetation coverage, and land cover factors in Fujian province, southeast China. Our results indicated that more forest fires occurred with lower elevation, flatter terrain, and higher population density areas. The global models showed that the precipitation and relative humidity had positive impacts on fire occurrence over the study area. In contrast, the GWR models revealed that the precipitation was positively correlated with fire occurrence in the west of Fujian, but negatively across the eastern coastal regions. The correlation between relative humidity and fire occurrence was also spatially different, positive in the north and negative in the center of Fujian. There existed overdispersion and spatial non-stationarity in the forest fire count data. The assessment of model fitting and predictions showed that GWNBR fit the forest fire count data better than other models, produced more precise and stable model parameter estimation, and yielded more realistic spatial distributions of model predictions. Thus, GWNBR is an effective and appropriate method for analyzing the occurrence of spatially varying and over-dispersed forest fires, and could provide better insight into forest fire mapping, prevention, and management based on local character prospects.

Keywords: geographically weighted Poisson regression; geographically weighted negative binomial regression; forest fire count; overdispersion.

1. Introduction

Forest fires spread across a large area of forested land and lead to significant, long-lasting impacts on ecological, social, and economic systems (Scott *et al.* 2013). In recent decades, forest fire occurrence has risen, along with the severity and damage they have inflicted. Canada has averaged 6863 fires and 2.6 million hectares burnt annually during 2006-2015, and the average annual national fire management expenditures exceeded CAD $800 million (Stocks and Martell 2016). There were 71,499 wildfires in the USA in 2017, burning about 4.05 million hectares, higher than the 10-year average (National Interagency Fire Center 2017). In China, about 10,000 wildfires occur with approximately 820,000 hectares burnt area each year (Guo *et al.* 2015).

Many models and methods have been developed to understand the influencing factors on fire occurrence and predict their potential threats to the environment, property, and lives (Costafreda-Aumedes *et al.* 2018). In recent years, advances in computer technology and software have led to wide applications of regression methods in forest fire modeling. Regression techniques provide the relationships and predictions between forest fires and explanatory variables such as vegetation patterns, fuel moisture conditions, meteorological variables, and some socio-economic factors (e.g., population density, GDP, education level, etc.). Logistic regression and multiple linear regression are two widely used methods on the study of fire ignition probability and driving factor analysis in different regions of the world (Vega Garcia *et al.* 1995; Andrews and Bradshaw 1997; Preisler *et al*. 2004; Brillinger *et al.* 2006; Guo *et al.* 2016a; Syphard *et al.* 2007; Oliveira *et al.* 2012). Both methods apply a square-root transformation to change the number of fires to a continuous scale. Sebastián-López *et al.* (2008) modeled ten-year fire danger based on a multiple linear regression, in which the dependent variable (the annual average of fire occurrence) was transformed using a natural logarithm.

On the other hand, instead of transforming discrete count data to a continuous scale, the expected number of fires can be estimated via generalized linear models. Poisson model has been considered a reliable method for fire occurrence modeling and risk analysis since the 1970s (Cunningham and Martell 1973; Dayananda 1977; Gill *et al*. 1987; Mandallaz and Ye 1997; Wotton *et al.* 2003, 2010). In addition to Poisson regression, negative binomial (NB) regression has also been applied in fire modeling studies, which have indicated that NB regression performs better in model fitting and prediction when the fire data (dependent variable) is dispersed (Cardille *et al.* 2001; Xiao *et al*. 2011; Chas-Amil *et al.* 2015).

The above models can be classified as global models, meaning that a single model with one set of model parameters can be used to explain the entire study region. However, spatial autocorrelation and heterogeneity exist across forest ecosystems, and the relationships between forest fires and environmental factors are nonstationary across space. Geographically weighted regression (GWR) was introduced as an alternative approach for modeling the spatially nonstationary fires across locations by Fotheringham *et al.* (2002). This approach considers varying relationships in space and allows local variations to be taken into account. With geographical information system (GIS) technology, spatially varying regression coefficients of GWR models can be visualized to identify local trends and spatial "hotspots" (Fotheringham and Brunsdon 2010). Further, GWR methodology has been extended to the framework of generalized linear models, such as geographically weighted logistic regression (GWLR) and geographically weighted Poisson regression (GWPR) (Nakaya *et al.* 2005). These developments have provided an appropriate foundation for modeling spatially varying binary and count response variables in the field of forest fires (Koutsias *et al.* 2010; Rodrigues *et al.* 2014; Guo *et al.* 2016b).

However, GWPR is more challenging than a global Poisson model, due to a common problem of overdispersion in spatial count data (Haining *et al.* 2009). Overdispersion is concerned with the strict requirement of a Poisson distribution for the count response variable to exhibit equal mean and variance. Given the nature of rare events, the occurrence (count data) of forest fires usually has much larger variance than the mean, because zero counts tend to occur more often than higher numbers of fire occurrence. This is particularly true for spatially clustered data such as the count of events in a census tract due to spatial heterogeneity within and between small geographical areas. If overdispersion is ignored, model fitting will underestimate the standard errors for Poisson regression model coefficients and lead to biased hypothesis testing (Lee 2011).

To model spatial count data with overdispersion, it may be more appropriate to use a negative binomial distribution instead of a Poisson distribution. Da Silva and Rodrigues (2014) proposed the geographically weighted negative binomial regression (GWNBR) method for incorporating spatial count data with overdispersion. Including spatial effects into statistical models is valuable for understanding relationships geographically and identifying local "hot spots" of high fire risks. In this study, we applied two global models (i.e., Poisson and Negative Binomial) and two GWR models (i.e., GWPR and GWNB) to develop forest fire prediction models and identify the driving factors of forest fire occurrence in a sub-tropical region of China. The results from each individual model were compared and the localized significant explanatory variables for prediction and prevention of forest fires were targeted. This study can improve the comprehensive understanding of the applicability of global and GWR models on forest fire studies.

2. Data and Methods

2.1 Study area

Fujian province is located in a sub-tropical region of China with a total land area of 124,000 km^2 (Figure 2.1). It is ranked the highest forest coverage in the nation (about 66% of Fujian province is covered by forests and vegetation), but experiences high annual forest fire incidences, with nearly 15,000 forest fires occurring from 2000 to 2010 (Guo *et al*. 2017). The dominant tree species in the province include Massoniana (*Pinus massoniana* Lamb.), Chinese fir (*Cunninghamia lanceolate* (Lamb.) Hook), Casuarina (*Casuarina equisetifolia* L.), Pubescens (*Phyllostachys heterocycle* (Carr.) Mitford cv. *Pubescens*), and others. The climate is warm and humid with an average annual rainfall of 1400 – 2000 mm and average temperature of 17 – 21 °C. Forest fire season typically spans from September to April (Guo *et al.* 2016a).

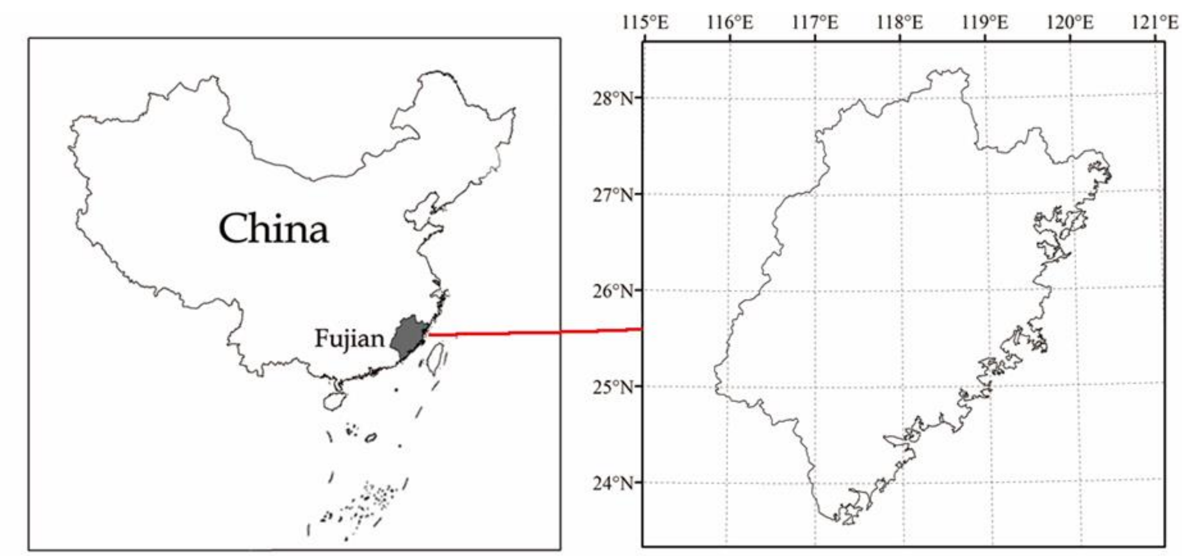


Figure 2.1 Location map of Fujian province, P.R. China.

2.2 Data preparation

2.2.1 Fire data (response variable)

In this study we used Medium Resolution Imaging Spectroscopy (MODIS) to record the spatial distribution of fire pixels in Fujian province from 2001 to 2016. This product is considered a reliable and suitable source for monitoring forest fires (Justice *et al*. 2002). We obtained a daily forest fire product (MOD14A1) with a resolution of 1 km, which has been widely used in recent forest fire studies (Amraoui *et al*. 2015; Guo *et al*. 2017). Since this product cannot distinguish forest fires from non-forest fires that occur in cities/towns, construction sites, agricultural lands, and other areas, we further processed the fire data by: (1) removing the fire points in cities/towns, construction sites, and farmland based on a 1 km resolution land-use map (the map is provided by Resource and Environmental Data Cloud Platform (http://www.resdc.cn/Default.aspx) and (2) extracting fire points based on the time of fire occurrence within the fire season (September 15 to April 30 of the following year). All forest fire points were recorded using geographical coordinates. Create Fishnet and Spatial Join in ArcGIS 10.2 (ESRI 2010) were then used to divide Fujian province into 4 × 4 km grids (a total of 7433 grids). Thus, the response variable was the total number of forest fire occurrences in each grid during a period of 16 years (2001 – 2016). Figure 2.2 shows the frequency distribution of the forest fire occurrence.

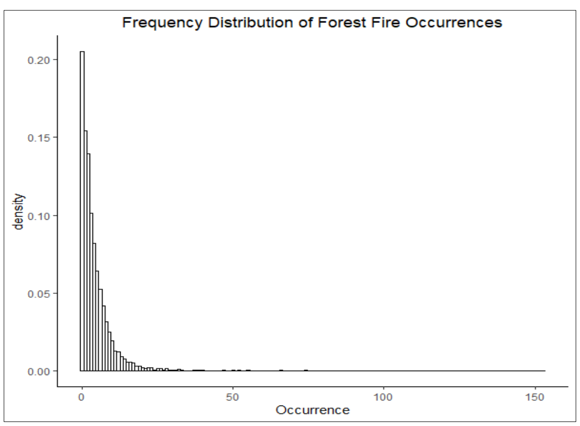


Figure 2.2 Frequency distribution of forest fire occurrences.

2.2.2 Potential driving factors (explanatory variables)

A total of 15 explanatory variables were collected in this study.

Topography

Topographic variables included elevation (km), slope (degree), and aspect index. High resolution (25 m) Digital Elevation Model (DEM) data was collected from the National Administration of Surveying, Mapping and Geoinformation of China (http://www.gscloud.cn/sources). Using 3D Analysis in ArcGIS 10.2, the slope and slope direction were derived from DEM, and aspect was then converted into an aspect index using the following formula (Guo *et al.* 2017): Aspect index = $\cos(\theta \times \pi / 180)$, where θ is the degree of slope generated in ArcGIS ranging from 0 – 360° so that the aspect index ranges from -1 to 1. An aspect index value closer to 1 indicates higher potential solar radiation. Average elevation, slope, and aspect index were then extracted using Zonal Statistics as Table in ArcGIS 10.2 for each grid.

Meteorology

Meteorological variables included precipitation (mm/day), temperature (°C), and relative humidity (%), which were derived from the HadCM2 global climate model (Guo *et al.* 2016b), an ocean-air coupling model developed by the Hadley Centre. ArcGIS 10.2 was used to calculate annual average daily values of each meteorological variable, and then Zonal Statistics as Table was used to extract averages of each meteorological variable for each grid. Precipitation and temperature impact the occurrence of forest fires by limiting the fuel moisture content. Therefore, they are reasonable and effective alternative fuel factors when other fuel factors are not available. In addition, the annual meteorological factor is a traditional and better indicator to measure the influence of climate change on forest fires, compared to the average meteorological data during the period of forest fire (Scholze *et al*. 2006; McCoy and Burn 2005; Xystrakis *et al.* 2013).

Human factors

Human factors included the socioeconomic variables (per capita GDP and population density) and infrastructural variables. GDP and population data was obtained from Resource and Environmental Data Cloud Platform (http://www.resdc.cn/Default.aspx). Data included grid population density and per capita GDP for the years 2000, 2005, 2010, and 2015 at 1 km resolution. Based on raster population and GDP data, the Raster Calculator tool in ArcGIS10.2 was used to calculate average annual growth rates of population and per capita GDP from 2000 to 2015. Average population and per capita GDP from 2000 to 2015 were then extracted using the Zonal Statistics as Table for each grid. Infrastructural variables included road density (km/km^2, ratio of road length to the grid area) and water density. An 1:250,000 vector map of infrastructure is provided by the National Geomatics Center of China (http://www.ngcc.cn/), and ArcGIS10.2 were used to evaluate the ratio of the length of road and area of water within each grid to the grid area.

Vegetation coverage and land coverage

Vegetation coverage is used to indicate the total amount of live and dead fuels above the surface. One practical estimation method uses the Normalized Difference Vegetation Index (NDVI). NDVI data was derived from the MODIS NDVI product with a spatial resolution of 500 m, provided by the Geospatial Data Cloud (http://www.gscloud.cn/). Variables of land use features were estimated from the Resource and Environmental Data Cloud Platform in China (http://www.resdc.cn/Default.aspx). It is a 1 km resolution raster data, providing the spatial distribution of vegetation types by digitizing the collections of vegetation type in China on the scale 1 : 1 million. Forest, including subtropical evergreen broad-leave forest type and mixed conifer and broad-leave forest type, covers about 64.95% of the total area. Vegetation type of shrub contains the subtropical evergreen broad-leave shrub, tropical evergreen broad-leaved shrub, and deciduous and broad-leaved shrub, taking about 20.40% of the study region. The cultivated land, fruit forest, and non-timber product forest are categorized into the cropland, covering about 12.6% of the total area. Grass (subtropical grass and tropical grass) only covers 1.69% of the total area. About 0.36% of the total area is the development land. According to the survey, regional forest coverage proportion is 66.80% and the agriculture land is about 11%, indicating the land classification of the raster data is close to reality. ArcGIS10.2 was used to calculate the proportion of each land cover in each grid.

2.2.3 Preliminary selection of variables

A multicollinearity analysis was performed before model fitting. The variance inflation factor (VIF) was used to detect the multicollinearity problems among the explanatory variables. In general, a VIF above 10 indicates that the parameter estimation and its standard error of an explanatory variable are impacted and damaged by multicollinearity. We used global Poisson model to test the multicollinearity, and resulted in the elimination of the socioeconomic variable of GDP because its VIF was 18.56. Therefore, there were a total of 14 explanatory variables used to fit the models in this study. Summary of fire occurrence (response variable) and 14 predictor variables were listed in Table 2.1. Figure 2.3(a) illustrated the spatial distribution of forest fire points, and Figures 2.3(b) – (p) presented the spatial distribution of predictor variables.

Table 2.1 Descriptive statistics of response and predictor variables.

Variable	Mean	Median	Std Dev	Minimum	Maximum
Fire Occurrence	4.2500	3.000	6.4271	0.0000	153.000
Elevation (km)	0.4885	0.4688	0.2714	-0.0064	1.7577
Slope (degree)	19.78	20.72	5.77	0.105	37.68
Aspect Index	-0.0073	-0.0116	0.1186	-0.5728	0.5523
Precipitation (mm/day)	3.9514	4.0659	0.5058	2.3232	4.9986
Temperature (°C)	21.9846	21.9041	1.5452	15.7449	25.3113
Humidity (%)	83.7273	83.7565	0.8788	81.3790	87.0881
Road Density	0.3632	0.3600	0.0804	0.00002	0.8427
Water Density	0.5969	0.5584	0.1962	0.0015	1.6537
Population (1000 people)	0.2839	0.1335	0.2714	0.0568	17.8995
NDVI	0.7742	0.7910	0.0662	0.3287	0.8729
Forest Cover (%)	51.752	55.374	39.487	0.000	100.00
Shrub Cover (%)	18.608	0.000	28.160	0.000	99.99
Grass Cover (%)	1.409	0.000	9.583	0.000	99.99
Crop Cover (%)	10.161	0.000	22.499	0.000	99.99

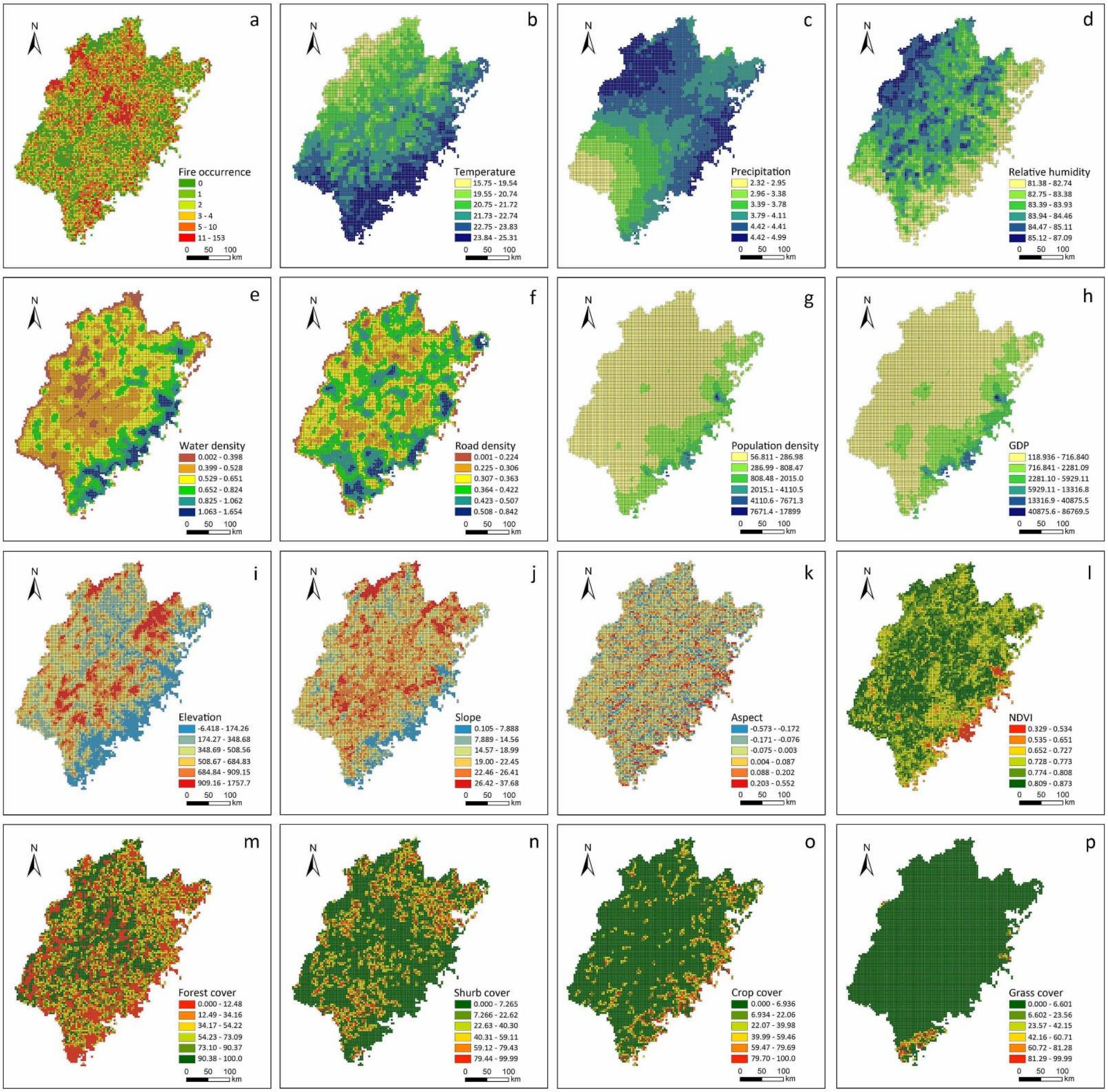


Figure 2.3 Spatial distribution of (a) Observed forest fire count, (b) Temperature, (c) Precipitation, (d) Relative humidity, (e) Water density, (f) Road density, (g) Population density, (h) GDP, (i) Elevation, (j) Slope, (k) Aspect index, (l) NDVI, (m) Forest cover, (n) Shrub cover, (o) Crop cover, and (p) Grass cover.

2.3 Models

2.3.1 Poisson regression

A count response variable (Y_i) has non-negative integer values (i = 1, 2,…, n), and Poisson regression is often used to model a count response variable Y_i against the underlying predictor variables (McCulloch and Searle 2001; Myers *et al*. 2002). The probability density function (*pdf*) of the Poisson distribution is:

$$P(Y_i) = \frac{e^{-\mu_i} \cdot \mu_i^{Y_i}}{Y_i\,!} \qquad [1]$$

where $P(Y_i)$ is the probability that the number of event (Y_i) occurred during a given time period, and μ_i is the parameter representing the expected value of Y_i. The Poisson distribution assumes equal mean and variance such that $E(Y_i) = \mu_i$ and $Var(Y_i) = \mu_i$. The explanatory variables are linked to the expected value μ_i via a link function such as a natural logarithm:

$$\ln(\mu_i) = \beta_0 + X_i\beta \qquad [2]$$

where X_i represents the explanatory variables, β_0 is the intercept coefficient, and β is the vector of the model slope coefficients. Thus, the expected value of μ_i can be predicted by the inverse link function $\hat{\mu}_i = e^{(\hat{\beta}_0 + X_i\hat{\beta})}$.

2.3.2 Negative binomial (NB) regression

Although Poisson regression is a common choice for modeling a count response variable, it is often criticized for its restrictive assumption of equal mean and variance. It is well known that the potential drawback of Poisson regression is the underestimation of standard errors of the model coefficients due to the overdispersion problem in the data. One way of dealing with this issue is to rescale the standard errors by the estimated dispersion parameter, while keeping the model coefficients unchanged. However, a better alternative for correcting the overdispersion problem is to use the negative binomial (NB) regression, which automatically builds in a dispersion parameter in its distribution function so that the estimation of both model coefficients and standard errors are corrected for the overdispersion in the data (McCulloch and Searle 2001; Myers *et al*. 2002). The unconditional distribution of Y_i can be written as:

$$P\left(Y_i\right)=\frac{\Gamma\left(Y_i+\frac{1}{\kappa}\right)}{\Gamma\left(Y_i+1\right)\Gamma\left(\frac{1}{\kappa}\right)}\left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa}\left(\frac{\kappa\mu}{1+\kappa\mu}\right)^{Y_i} \qquad [3]$$

where Γ denotes the gamma function, κ is the dispersion parameter, and the mean and variance of Y_i are:

$$E\left(Y_i\right)=\mu \qquad [4]$$

$$V\left(Y_i\right)=\mu+\kappa\mu^r=\mu\left(1+\kappa\mu\right) \qquad [5]$$

Thus, the Poisson model is the limiting model of the negative binomial model when $\kappa\rightarrow 0$. The common link function for the negative binomial regression is the same as the Poisson regression (Eq. [2]).

2.3.3 Geographically weighted regression (GWR)

To investigate the spatial variation or heterogeneity of a regression relationship, the data must be collected with the location coordinates (v_{xi}, v_{yi}) for each observation *i*. This local information allows for estimation of the localized regression coefficients of the relationship of interest. When GWR was first developed, the Gaussian assumption was assumed for the model error term (Fotheringham *et al.* 1998), expressed as:

$$Y_i=\beta_0\left(v_{xi},v_{yi}\right)+\sum_{k=1}^{p}\beta_k\left(v_{xi},v_{yi}\right)X_{ki}+\varepsilon_i \qquad [6]$$

where Y_i is the response variable, X_k is a set of p explanatory variables (k = 1, 2, …, p), $\beta_0(v_{xi}, v_{yi})$, $\beta_1(v_{xi}, v_{yi})$, …, $\beta_p(v_{xi}, v_{yi})$ are the regression coefficients for the kth predictor variable at the ith location, and ε_i is the random error term whose distribution is assumed $N(0, \sigma^2 I)$ with I denoting an identity matrix. The aim of GWR is to obtain the estimates of these functions for each predictor variable and each geographic location i. The estimation procedure is as follows: (i) draw a circle of a given radius around one particular location i (the center), (ii) compute a weight (w_{ij}) for each neighboring observation j according to the distance d_{ij} between location j and center i, and (iii) estimate the model coefficients using weighted least-square regression such that:

$$\hat{\beta}_i = \left(X' W_i X\right)^{-1} X' W_i Y \qquad [7]$$

where the weight matrix W_i is:

$$W_i = \begin{pmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{pmatrix} \qquad [8]$$

The weighting function is defined by the kernel type and size of kernel (bandwidth), which determines the geographical weight of the jth neighboring observation at the ith regression point. The weight should decrease gradually as the distance between i and j increases, until it reaches a constant or zero. The model parameter estimates are highly related to the kernel size, so the choice of the kernel is important in the modeling process of GWR.

2.3.4 Geographically weighted Poisson regression (GWPR)

The GWPR model is developed by adding geographical location into the standard Poisson regression. It uses a similar link function to Eq. [2] and is in the following form:

$$\ln\left(u_i\right) = \beta_0\left(v_{xi}, v_{yi}\right) + \sum_{k=1}^{p} \beta_k\left(v_{xi}, v_{yi}\right) X_{ki} \qquad [9]$$

where β_0 and β_k are the GWPR model parameters specifically describing the location of *i* with *x* and *y* coordinates.

2.3.5 Geographically weighted negative binomial regression (GWNBR)

GWNBR is an extension of the global model of NB regression that allows the spatial variation of parameters β_0, β_k, and κ. This local model is described as:

$$\ln(u_i) = \left[\beta_0(v_{xi}, v_{yi}) + \sum_{k=1}^{p} \beta_k(v_{xi}, v_{yi}) X_{ki}, \kappa_i(v_{xi}, v_{yi})\right] \qquad [10]$$

where β_0 and β_k are the GWPR model parameters specifically describing the location of *i* with *x* and *y* coordinates, and κ_i is the local dispersion parameter.

In this study, we applied the same Gaussian kernel function and bandwidths to both GWPR and GWNBR models. It is known that the bandwidth has profound impacts on model fitting, the spatial distribution of model predictions, and localized model coefficients (Fotheringham *et al*. 2002; Guo *et al*. 2008). If different bandwidths were used for GWPR and GWNBR, the modeling results would be less compatible between the two models because we would not be able to distinguish if the model differences were due to the models *per se* or the bandwidths used.

2.4 Model evaluation and comparison

Overdispersion in the response variable is always a concern when modeling count data. One way to detect the problem is to divide the model deviance, which measures the discrepancy between the observed and fitted response variable, by its degrees of freedom (df). If this deviance/df ratio is close to 1, there is no concern on the overdispersion; if it is greater than 1, an overdispersion problem may exist and some correction may be necessary (Myers *et al*. 2002).

In a GWR model, it is necessary to decide on an optimal bandwidth for model fitting (Fotheringham *et al.* 2002). There are three common ways of choosing the bandwidth: (1)

subjective choice, (2) based on the smallest cross-validation error, or (3) based on the smallest Akaike Information Criterion (AIC) (Fotheringham *et al.* 2002; Guo *et al.* 2008). In this study, we used AIC to decide the optimal bandwidth and related kernel function for estimating each GWR model. A number of variogram models (Bailey and Gatrell 1995) were used for the spatial data in order to find the optimal bandwidth and kernel function, resulting in a Spherical due to its smallest AIC. The estimated bandwidth was selected at 112,410 meters, depending on the residuals of global models.

To evaluate spatial variation or heterogeneity in the model coefficients of GWPR and GWNBR, we followed the approach in Chen *et al.* (2012). The interquartile range (IQR) of the coefficient estimates computed by the GWR localized models was compared to the standard error of the global estimates derived with a traditional regression model. When IQR is twice as large as the standard error, it indicates that spatial non-stationary exists in the relationships between the response variable and its accompanying predictor variables.

Model fitting was evaluated using AIC and mean squared errors (MSE) (Burnham and Anderson 2004). Smaller AIC or MSE imply better model fitting performance. To evaluate the spatial autocorrelation of the residuals, the Geary's contiguity ratio (Geary's C) was calculated. The closer to 1 the Geary's C, the lower the spatial dependence of the residual will be, and hence, the model accounts for more spatial structure problems. Chi-square (χ^2) goodness of fit was used to compare model prediction performances (Terceiro 2003; Zhen *et al*. 2018). In addition, the predicted fire occurrence of the four models and the spatial distribution of model coefficient for each explanatory variable were mapped using ArcGIS10.2.

3. Results

3.1 Model fitting and prediction accuracy comparison

3.1.1 Overall comparison between global and GWR models

The statistics of model fitting are listed in Table 2.2, including AIC, MSE, and Geary's C of model residuals with corresponding Z-test and *p*-value. and NB models, The AIC values were 53282.4 for the global Poisson model and 36446.6 for the global NB model, which were much greater than those of both GWPR (27066.3) and GWNBR (22886.7) models. The MSE of the GWR models (28.4 for GWPR and 31.3 for GWNBR) were smaller than those of the two global models (36.2 for Poisson and 36.7 for NB). Thus, it was evident that the GWR models fitted the fire occurrence data better than the two global models. Additionally, the Geary's C of the two global model residuals was significantly smaller than one, indicating a clustered spatial pattern (i.e., positive spatial autocorrelations among the model residuals). In contrast, the Geary's C of the GWNBR model residuals was not significantly different from one, representing a desirable random spatial pattern; while the Geary's C of the GWPR model residuals was significantly larger than one, indicating a dispersed / uniform spatial pattern (i.e., negative spatial autocorrelations among the model residuals) (Table 2.2).

Table 2.2 Statistics of model fitting and prediction for global and GWR models.

Statistics	**Global Poisson**	**GWPR**	**Global Negative Binomial**	**GWNBR**
AIC	53282.4	27066.3	36446.6	22886.7
MSE	36.21	28.38	36.68	31.32
Geary's C	0.9918	1.0090	0.9888	1.0035
Z-score	-3.63	3.99	-4.98	1.56
p-value	0.0003	<.0001	<.0001	0.1187
Goodness of fit (χ^2)	5234.6	3949.6	5157.9	4129.1
Deviance / df	4.6593	-	1.1117	-

In terms of model predictions, the GWR models were closer to the observed forest fire counts than the global models. The predicted counts from the global models did not exceed 20, indicating that no grid across Fujian province had fire occur more than 20 times during this time period (Figures 2.4(a) and (c)). In contrast, the predictions from the GWR models showed wider ranges, where some locations were predicted greater than 20 fire occurrences during this time period (Figures 2.4(b) and (d)). Regarding the spatial distribution of model predictions, the GWR models showed that high incidences (> 10) of forest fires were mainly concentrated in the northwest and center regions, while only few high incidences of forest fires predicted by the global models randomly scattered across the study area (Figures 4(a) and (c)). To compare the model predictions for GWR, we used Chi-square (χ^2) goodness of fit statistics (Table 2.2). In this study, the response variable (i.e., forest fire counts), is divided into 17 categories, which are 0, 1, 2…, 15, and > 15. We then compared the predicted with observed counts respectively. The Chi-square statistics of the GWR models were smaller than the global models, indicating that GWPR and GWNBR were better than Poisson and NB in model prediction performance.

3.1.2 Comparison between global Poisson and NB models

In this study, the deviance/df ratio of the global Poisson model was 4.6593 (Table 2.2), showing the existence of the overdispersion problem. But the deviance/df ratio of the global NB model was 1.1117, close to 1. Therefore, the NB model was indeed a better choice than the Poisson model for handling the overdispersion in our fire count data. Furthermore, the AIC of the two models confirmed that the global NB model (36446.6) was superior to the global Poisson model (53282.4) for fitting the fire count data (Table 2.2).

The predictions from both global Poisson and NB models were between 0 and 20, suggesting that 0 or at most 20 fires were expected in each grid during this time period. Both global models projected that the areas with frequent forest fires were mainly concentrated in the northwest and southeast of the study area (Figures 2.4(a) and (c)).

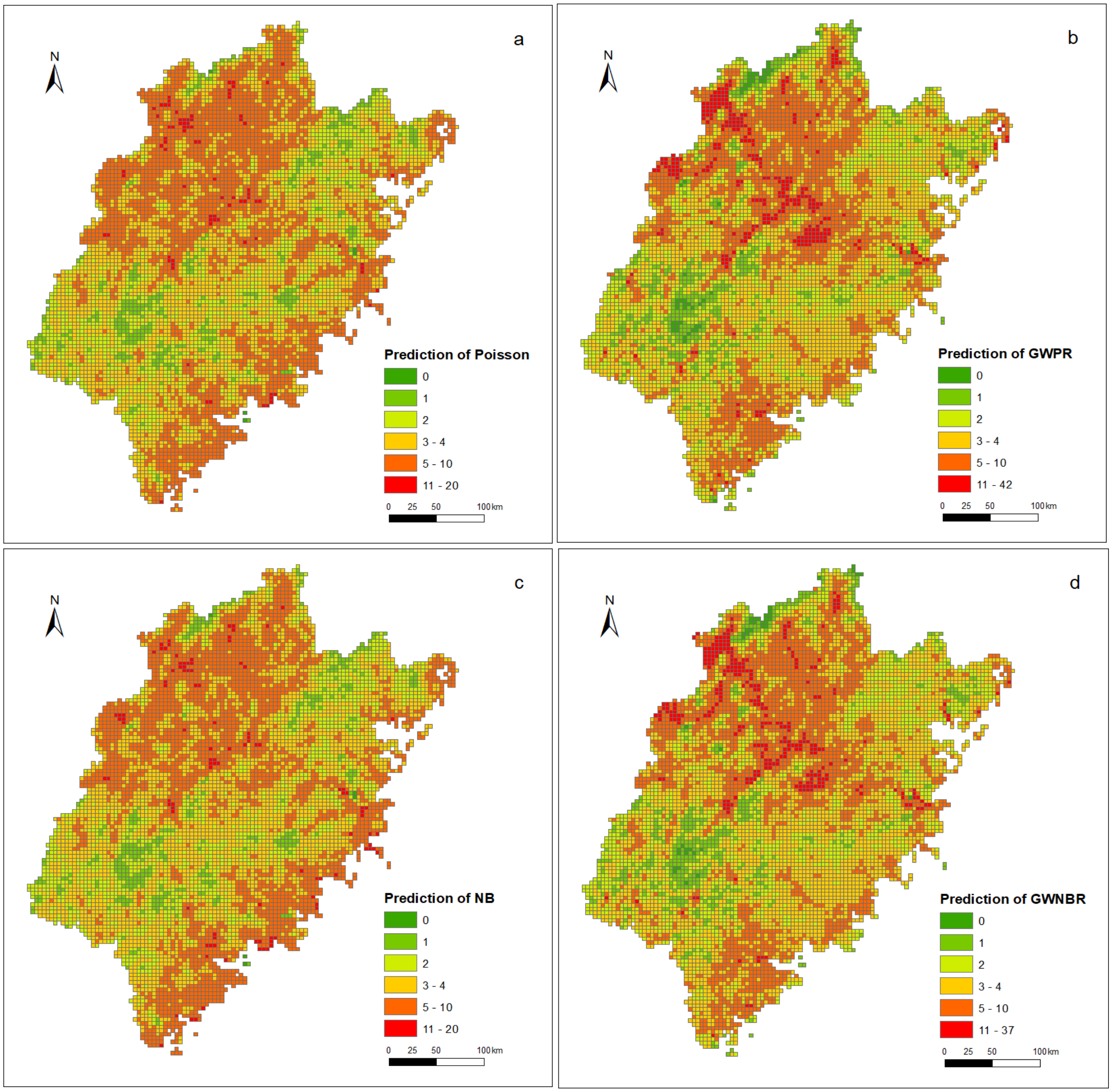


Figure 2.4 Spatial distributions of model predictions from (a) global Poisson, (b) GWPR, (c) global NB, and (d) GWNBR.

3.1.3 Comparison between GWPR and GWNBR models

Similar to the global models, the GWNBR model (22886.7) had a much smaller AIC than the GWPR model (27066.3). The Geary's C of GWNBR revealed that its model residuals were randomly distributed. The results of AIC and Geary's C implied that GWNBR was better than GWPR for fitting the fire count data. The spatial distributions of the fire predictions from the two

GWR models were similar, i.e., the areas with high frequencies of forest fires (i.e., predicted fire occurrence > 10) were concentrated in the northwest and center of Fujian province. The Chi-square statistics implied the GWPR was slightly better than GWPR in terms of model prediction (Table 2.2).

3.2 Variability of significant explanatory variables

3.2.1 Overall comparison between global models and GWR models

The relationship between forest fire occurrence and explanatory variables estimated by the global models is assumed constant and stationary across the study area. On the other hand, the model coefficients of the GWR models are spatially varied from location to location. The results of the GWR models showed that the relationship between fire occurrence and explanatory variables was spatially non-stationary. Compared to the global models, the GWR models highlighted the spatial heterogeneity of the relationships between response and explanatory variables. The significance level was chosen at $\alpha = 0.05$ for this study. Both global models showed that elevation, slope, aspect index, humidity, precipitation, population, NDVI, road density, water density, grass cover, and crop cover were significantly correlated with the forest fire counts (Table 2.3). However, the significance of the GWR model coefficients may not be consistent across the study area, but varied from location to location. Figures 2.5 and 2.6 demonstrated the locally significant model coefficients.

Table 2.3 Model parameter estimation of global Poisson and Negative Binomial models.

Parameters	**global Poisson**			**global Negative Binomial**		
	Estimate	**Standard Error**	***p*-value**	**Estimate**	**Standard Error**	***p*-value**
Intercept	-13.5939	1.4903	<.0001	-12.6004	3.2821	<.0001
Elevation	-2.000	0.0457	<.0001	-1.7442	0.0880	<.0001
Slope	0.00589	0.00203	0.0037	-0.0096	0.00483	0.0468
Aspect index	-0.2876	0.0482	<.0001	-0.3075	0.1086	0.0047
Humidity	0.2076	0.0157	<.0001	0.1840	0.0346	<.0001
Temperature	-0.01828	0.00964	0.0579	0.0145	0.0215	0.4989
Precipitation	0.2608	0.01767	<.0001	0.3032	0.0360	<.0001
Population	-0.5145	0.0199	<.0001	-0.5165	0.0387	<.0001
NDVI	-3.3388	0.1553	<.0001	-2.9133	0.3788	<.0001
Road density	2.0681	0.0788	<.0001	1.8798	0.1830	<.0001
Water density	-0.4232	0.0415	<.0001	-0.3460	0.0968	0.0004
Forest cover	0.0179	0.0171	0.2940	-0.01007	0.03778	0.7898
Shrub cover	-0.03079	0.0232	0.1844	-0.02973	0.05073	0.5578
Grass cover	0.3517	0.0560	<.0001	0. 2806	0. 1314	0.0328
Crop cover	-0.241	0.0291	<.0001	-0. 3511	0.07026	<.0001
Dispersion	-	-	-	0.9026	0.0204	-

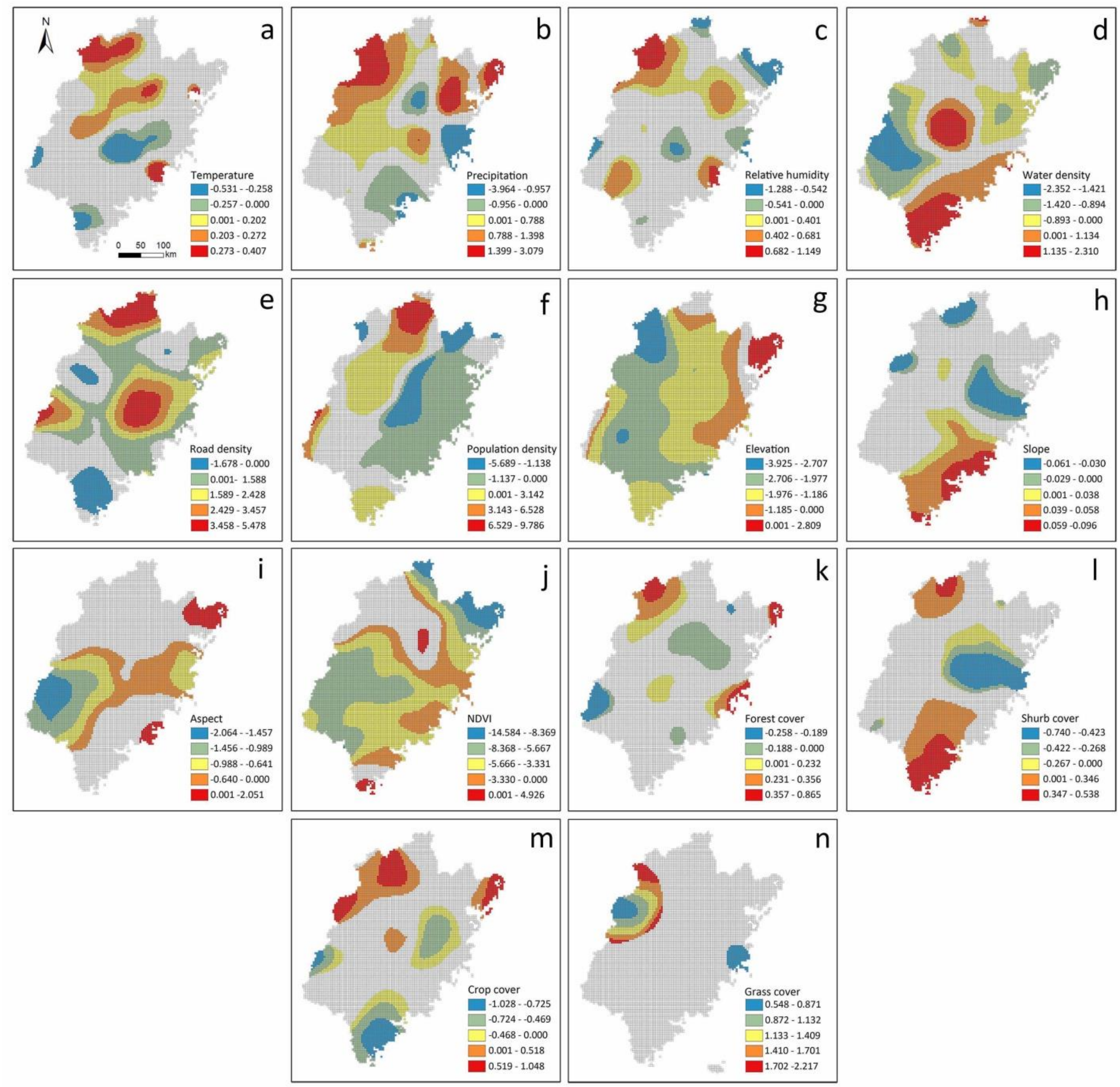


Figure 2.5 Spatial distribution of significant model coefficients of (a) Temperature, (b) Precipitation, (c) Relative humidity, (d) Water density, (e) Road density, (f) Population density, (g) Elevation, (h) Slope, (i) Aspect index, (j) NDVI, (k) Forest cover, (l) Shrub cover, (m) Crop cover, and (n) Grass cover of the GWPR model.

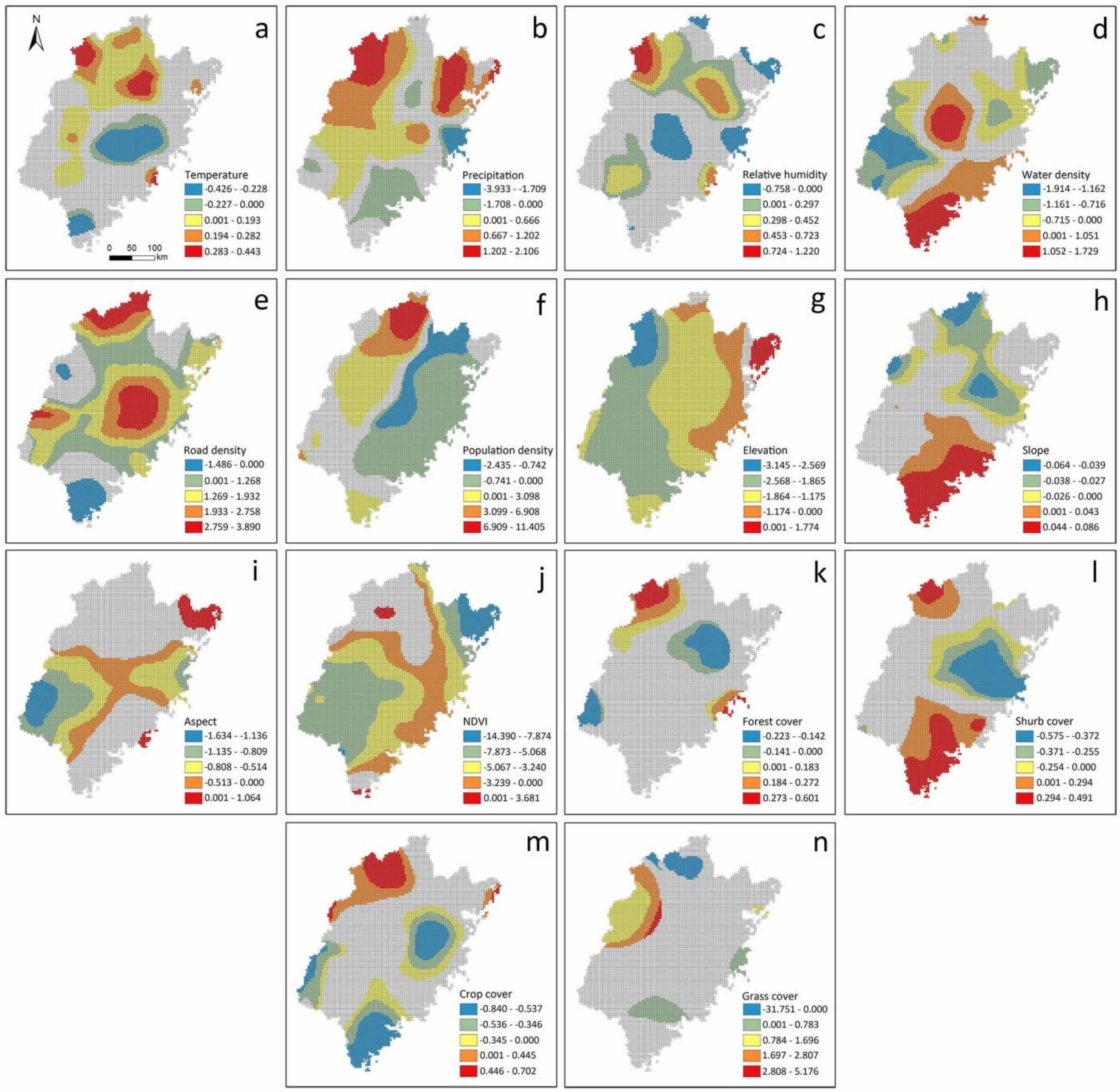


Figure 2.6 Spatial distribution of significant model coefficients of (a) Temperature, (b) Precipitation, (c) Relative humidity, (d) Water density, (e) Road density, (f) Population density, (g) Elevation, (h) Slope, (i) Aspect index, (j) NDVI, (k) Forest cover, (l) Shrub cover, (m) Crop cover, and (n) Grass cover of the GWNBR model.

3.2.2 Comparison of explanatory variables of global models

The estimated coefficients of the global Poisson and NB models are listed in Table 2.3. The relationship between significant explanatory variables and forest fire occurrence in the NB models had the same correlation as in the global Poisson model. Additionally, in both global Poisson and NB models, the estimated model coefficients of humidity, precipitation, road density, and grass cover were significantly positive, indicating that greater relative humidity and/or precipitation, denser road and grass cover may lead to more forest fires (see discussion section for details). In contrast, the estimated model coefficients of water density, population density, NDVI, crop cover, and two topographic factors (elevation and aspect index) were significantly negative, indicating that forest fires were less likely to occur in developed areas, relatively high elevations, less sunshine radiation and/or in the areas with low crop cover.

3.2.3 Comparison of spatial variability of significant variables in GWR models

The GWR Poisson and NB model coefficients of all fourteen explanatory variables were spatially varied because their IQRs were at least twice as large as the standard errors of the corresponding global model coefficients. Our results suggested that the relationships between forest fire occurrence and topographical, meteorological variables, human variables (e.g., population, water, and road density), and vegetation and land use cover were indeed heterogeneous across the study region (Table 2.4). The spatial distributions of the significant model coefficients of explanatory variables between GWPR and GWNBR were similar. Additionally, the coefficients of explanatory variables were only significant in some locations of the study area, and the influencing direction (the sign of coefficients) and power (estimated value of coefficients) were also spatially varied (Figures 2.5 and 2.6). For example, the temperature was positively correlated to fire occurrence in the north, but negatively correlated to fire counts in the center of the province. Similarly, the precipitation and population density showed negative relationships with forest fires in the east and southeast of Fujian, but positive relationships in the north.

Table 2.4 Summary of model parameter estimation of geographically weighted Poisson and Negative Binomial models.

Statistics	Model	$\beta_{Intercept}$	$\beta_{Elevation}$	β_{Slope}	$\beta_{Aspect\ index}$	$\beta_{Humidity}$	$\beta_{Temperature}$	$\beta_{Precipitation}$	$\beta_{population}$	β_{NDVI}	$\beta_{Road\ density}$	$\beta_{Water\ density}$	$\beta_{Forest\ cover}$	$\beta_{Shrub\ cover}$	$\beta_{Grass\ cover}$	$\beta_{Crop\ cover}$	Local dispersion
Mean	GWPR	-4.292	-1.617	0.00569	-0.3442	0.0829	0.0286	0.3232	0.6032	-3.966	1.3094	-0.1415	0.0252	-0.0137	-5.6522	-0.0375	-
	GWNBR	-5.846	-1.628	0.00003	-0.3219	0.1005	0.0291	0.3119	0.7773	-3.645	1.3393	-0.0603	0.0107	-0.0233	-5.7854	-0.1058	0.0565
Median	GWPR	-5.659	-1.760	0.00227	-0.2787	0.1022	0.0501	0.3207	-0.1382	-3.880	1.2280	-0.2621	0.0204	-0.0115	0.0000	0.0001	-
	GWNBR	-5.902	-1.707	-0.00516	-0.2473	0.1099	0.0507	0.2918	-0.1055	-3.667	1.3831	-0.1507	-0.0033	-0.0278	0.0000	-0.0995	0.0510
Min	GWPR	-108.98	-3.925	-0.0612	-2.0639	-1.2875	-0.5312	-3.9636	-5.6887	-14.584	-1.6778	-2.3525	-0.2585	-0.7401	-428.52	-1.0281	-
	GWNBR	-118.33	-3.415	-0.00643	-1.6342	-0.7575	-0.4262	-3.9332	-2.4348	-14.390	-1.3371	-1.9136	-0.2233	-0.5752	-1234.5	-0.8397	0.0190
Max	GWPR	111.264	2.809	0.0956	2.0505	1.1494	0.4068	3.0787	9.7859	4.926	5.4778	2.3102	0.8653	0.5377	301.45	1.0482	-
	GWNBR	78.395	1.774	0.0863	1.0643	1.2201	0.4429	2.1058	11.4046	3.681	3.5797	1.7295	0.6013	0.4911	411.82	0.7017	0.1422
IQR	GWPR	38.5508	0.9821	0.0313	0.6457	0.3947	0.2537	1.0017	1.3115	4.0899	1.7066	1.3304	0.2282	0.4003	3.2728	0.4229	-
	GWNBR	31.5258	0.8528	0.0371	0.5888	0.3247	0.2188	0.9027	1.3347	3.6160	1.3479	1.0937	0.1580	0.3502	1.8188	0.4297	0.0248
Std	GWPR	33.6224	0.9470	0.0286	0.5800	0.3486	0.1795	0.8532	2.2621	2.7867	1.4283	0.9207	0.1548	0.2734	29.089	0.3684	-
	GWNBR	28.3133	0.7823	0.0300	0.4405	0.2816	0.1659	0.7859	2.3213	2.8930	1.1447	0.7916	0.1191	0.2484	45.451	0.3308	0.0217
Ste†·	GWPR	1.4903	0.0457	0.00203	0.0482	0.0157	0.00964	0.01767	0.0199	0.1553	0.0788	0.0415	0.0171	0.0232	0.056	0.0291	-
	GWNBR	3.2821	0.088	0.00483	0.1086	0.0346	0.0215	0.036	0.0387	0.3788	0.183	0.0968	0.037	0.05	0. 1314	0. 0703	0.0204
Status	GWPR	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	-
	GWNBR	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	S

† Note: Standard error (Ste) was estimated from the global regression; NS indicates spatially nonstationary.

4. Discussion

4.1 Comparison of prediction models for forest fire occurrence

Our study indicated that the GWR models had better model fitting than the global models, which was indeed expected as the GWR models estimated the local parameters of each location and effectively explained the spatial variability of the response variable (Wu and Zhang 2013). Our results were in line with previous studies that reported better model fitting and predictions of GWR than global models (Guo *et al*. 2016b; Rodrigues *et al*. 2014).

GWPR produced less significant model coefficients of the explanatory variables such as water density, road density and forest cover than did GWNBR (Figures 2.5 and 2.6). A possible reason for these differences was that GWNBR estimated the dispersion parameter location by location. The spatial distribution of the local dispersion parameters is shown in Figure 2.7. It was evident that larger dispersion parameters were clustered along the southeast coast of Fujian province, where the forest fires occurred less than 2 or more than 10 times (Figure 2.3(a)). The local dispersion parameters then gradually decreased in the northwest areas. This was reasonable because either extreme small or great forest fire occurrences were detected in the southeast, which likely led to a greater overdispersion compared to other regions. The existence of varying overdispersion across the study area clearly suggested the need to correct the model standard errors location by location. Therefore, GWNBR is more appropriate than GWPR for modeling such dispersed fire count data.

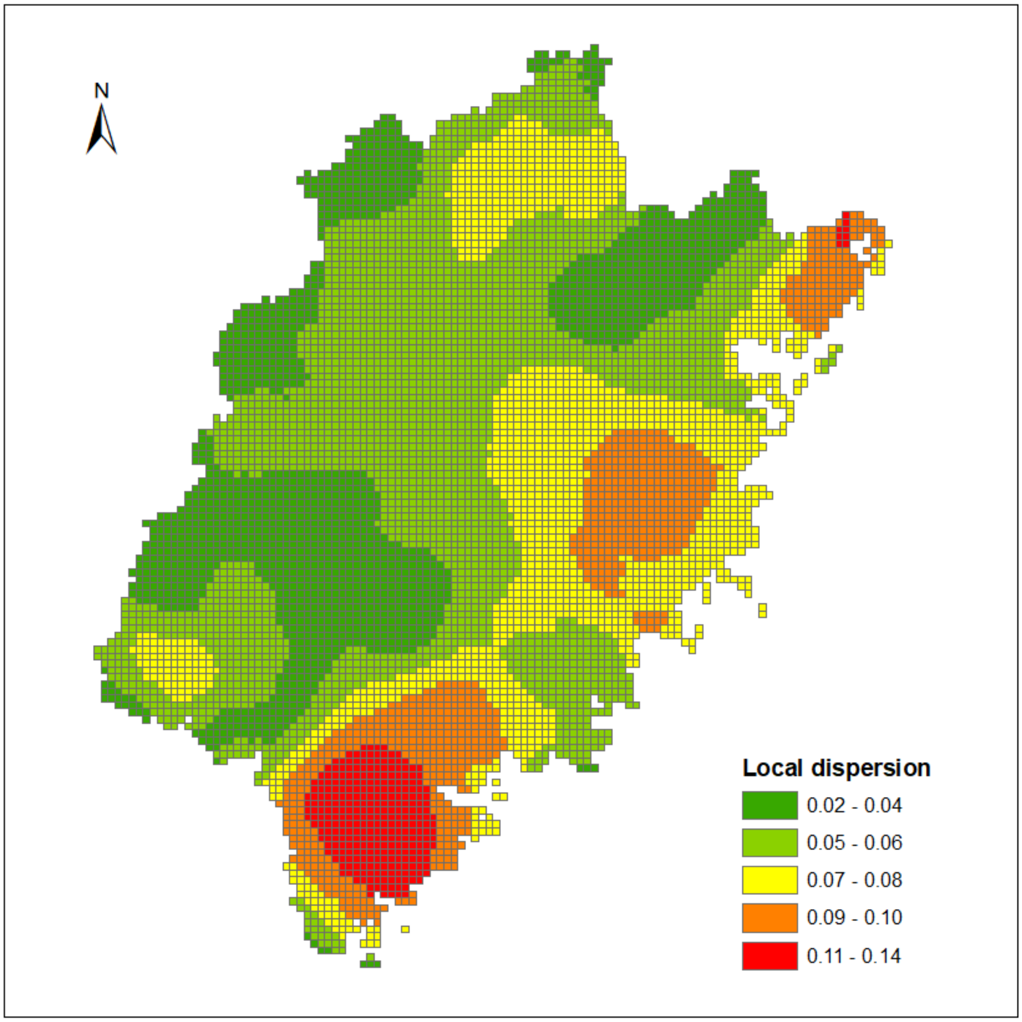


Figure 2.7 Spatial distribution of local dispersion parameters of GWNBR model.

4.2 Influence of drivers on forest fire occurrence in Fujian

The drivers and their influence on fire occurrence in Fujian varied between global and GWR models. Compared to the global models, the GWR models explained more specific spatial relationships between drivers and fire occurrence. Those drivers were not consistently positive or negative influencing the fire occurrence across the study area.

Based on the global models, our results indicated that the number of forest fire occurrences in Fujian was increased with lower elevation, flatter terrain, and denser road. The findings are generally supported by the GWR models. Human activities more likely occur in the low elevation, flatter terrain, and dense road nets, which may lead to higher fire risks. Several empirical studies on forest fires and their driving factors in Fujian province have found that human activities are more frequent at low-altitude, in areas with gentle slopes, and close to roads, which lead to a higher probability of human-caused fires (Guo *et al*. 2016a; Guo *et al*. 2016c). However, according to the global models, it was found that areas with high population density had fewer forest fires, where human activity was also very frequent. The finding was consistent with the GWR models that showed a negative correlation between population density and fire occurrence in the eastern coastal areas. High population density tend to be concentrated in cities or developed areas with developed industry and low forest coverage, where the combustible material is not as continuous as the vegetation in remote forests (Vega Garcia *et al*. 1995; Maingi and Henry 2007; Miranda *et al*. 2012; Guo *et al*. 2016a, 2016c).

In addition, both global and GWR models similarly indicated that the aspect index (value from -1 to 1) was negatively associated with fire occurrence. The smaller the aspect index was, the closer the region was to the south and southwest, with stronger sunshine, higher temperatures, and faster evaporation. These features lead to dryer fuel making it more likely for forest fires to occur (Hu 2005). Generally, a higher NDVI means higher fuel load and higher probability and frequency of forest fires, but Fujian province shows the opposite effect. One explanation for this is that vegetation cover affects wildfire by affecting the temperature of fine fuel on the underlying surface. In the areas with high vegetation coverage, the surface temperature is low, which makes it difficult for soil moisture to evaporate, leading to higher fuel moisture content, thus lower likelihood of burning (Huang *et al*. 2015).

The GWR models illustrated that the model coefficients of crop cover were positive in the north, but negative in the south. One possible explanation was that the agricultural land was mainly distributed in the southeastern part of Fujian province. The forest coverage rate in these areas was relatively low, so the number of forest fires was negatively correlated with the crop area. In contrast, the forest area in the northern region is large, and the farmland is mostly in the forest interaction zone. As the crop area increased, the frequency of agricultural production activities increased, which was likely to lead to more forest fires. Additionally, slope was negatively associated with the fire occurrence in the north and had an opposite impact in the south. It was because more grassland distributed in southern Fujian. When the slope is large, the grass become drier and may cause more fires. There are more forests in the north, and fewer human activities occur in high-slope areas, resulting in low human-caused fire source.

Meteorological factors have been found to be important driving factors of forest fires in Fujian (Guo *et al*. 2017), and this was confirmed by our research. According to the global models, the precipitation and relative humidity have positive impacts on fire occurrence over the study area. Although the findings seem controversial to the general understanding that as rainfall increases, fewer fire will occur. One explanation for this is more rainfall and humidity are beneficial to the growth of ground-cover vegetation, and increased amounts of surface fuel load which increases the risk of forest fire occurrence. In the Kruger National Park of South Africa, van Wilgen *et al.* (2000) observed a strong positive correlation between precipitation rates and fire activity. Spessa *et al.* (2005) and Randerson *et al.* (2005) also found a similar positive association between precipitation and fire activity in north Australia using different satellite data sets. In contrast, the GWR models provided more specific spatial relationships between meteorological factors and fire occurrence. Similar to the global models, the precipitation is positively correlated with fire occurrence in the west of Fujian, but negatively in the eastern coastal regions. Since there is plenty of rainfall in the coastal areas, the role of precipitation is more likely a limiting factor on fire occurrence rather than a promoting factor of fuel load, which will positively impact the fire occurrence. The correlation between relative humidity and fire occurrence is also spatially different, positive in the north and negative in the center of Fujian (Figures 2.5 and 2.6).

4.3 Other potential models to consider

There are other potential candidate models for dealing with the overdispersion problem in a count data such as a quasi-Poisson model, which is consider an intermediate model between Poisson and NB models. The main problem with quasi-Poisson is that there is no corresponding distribution or likelihood for the model, and hence some extremely useful statistical tests and fit measures (e.g., AIC, LR and etc.) are unavailable. Researchers made different decisions and comments on which model is appropriate to the over-dispersed data (Gardner *et al*. 1995; Power and Moser 1999; Potts and Elith 2006; Ver Hoef and Boveng 2007). In recent years, people have paid attention to the comparison and selection between quasi-Poisson and negative binomial (NB) models. Seyoum *et al.* (2016) proposed an approach for detecting which model is appropriate to the count data with overdispersion. They identified a cut-off point by equating the two variance functions of quasi-Poisson and NB models. Then, if the mean of count response variable is less than the cut-off point, the negative binomial model should be considered; while if the mean of the variable of interest is greater than the cut-off point, the quasi-Poisson model is more appropriate. For our fire count data, we fit both quasi-Poisson and NB models to obtain their dispersion parameters as follows: the quasi-Poisson dispersion parameter $\Phi = 6.3220$ and the negative binomial dispersion parameter $\theta = 0.9026$, so that the cut-off point = $(\Phi\text{-}1) / \theta$ = 5.8963. Since the mean of the forest fire count was 4.25 < cut-off point, the NB model was more appropriate and preferred to the quasi-Poisson model in this study.

5. Conclusion

In this study, the spatial varying relationships between forest fire occurrence and driving factors in Fujian province from 2001 - 2016 were evaluated using two global models (Poisson and negative binomial (NB)) and two geographically weighted generalized linear models (GWPR and GWNBR). Our results indicated that, compared to the global Poisson and NB models, the GWR generalized linear models had better performance in model fitting, predictions, and spatial distributions of model predictions, as well as detecting the impact hotspots of the predictor variables. The GWR generalized linear models can effectively incorporate spatial dependence and non-stationary relationships in the count response variable. Simultaneously, we compared the performance of GWPR and GWNBR in modeling spatial count data with overdispersion and found the estimated model coefficients of GWNBR were more precise and stable than those of GWPR.

In addition, we determined the drivers and spatial distribution of subtropical forest fires in Fujian province, China based on the above methods. Two types of models (global and GWR) similarly indicated that the number of forest fire occurrences in Fujian was increased with lower elevation, flatter terrain, and denser road. Compared to the global models, the GWR models can indicate more specific spatial relationships between drivers and fire occurrence. For example, the precipitation and population density had different impacts on fire occurrence in the coastal areas than other regions of the study area. In summary, GWNBR is an effective and appropriate method for analyzing the occurrence of spatially varied and over-dispersed forest fires, It clearly indicated the key fire drivers and their influences and can provide reliable insight into forest fire mapping, prevention, and management based on local character prospects.

6. References

Amraoui M., Pereira M.G., DaCamara C.C., Calado T.J., 2015. Atmospheric conditions associated with extreme fire activity in the Western Mediterranean region. *The Science of the Total Environment* 524-525: 32-39.

Andrews P.L., Bradshaw L.S., 1997. Fires: Fire information retrieval and evaluation system - A program for fire danger rating analysis. USDA Forest Service, Intermountain Research Station General Technical Report INT-367. 64 pp (Ogden, UT).

Bailey T.C., Gatrell A.C., 1995. Interactive spatial data analysis. Longman Scientific and Technical. Routledge.

Brillinger D.R., Preisler H.K., Benoit J.W., 2006. Probabilistic risk assessment for wildfires. *Environmetrics* 17(6): 623-633.

Burnham K.P., Anderson D.R., 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research* 33(2): 261-304.

Cardille J.A., Venture S.J., Turner M.G., 2001. Environmental and social factors influencing wildfires in the upper Midwest, United States. *Ecological Applications* 11(1): 111-127.

Chas-Amil M.L., Prestemon J.P., McClean C.J., Touza J., 2015. Human-ignited wildfire patterns and responses to policy shifts. *Applied Geography* 56: 164-176.

Costafreda-Aumedes S., Comas C., Vega-Garcia C., 2018. Human-caused fire occurrence modelling in perspective: a review. *International Journal of Wildland Fire* 26(12): 983-998.

Chen V.Y., Deng W.S., Yang T.C., Matthews S.A., 2012. Geographically weighted quantile regression (GWQR): An application to mortality data. *Geographical Analysis* 44(2): 134-150.

Cunningham A.A., Martell D.L., 1973. A stochastic model for the occurrence of man-caused forest fires. *Canadian Journal of Forest Research* 3: 282-287.

Da Silva A.R., Rodrigues T.C.V., 2014. Geographically weighted negative binomial regression-incorporating overdispersion. *Statistics and Computing* 24(5): 769-783.

Dayananda P.W.A., 1977. Stochastic models for forest fires. *Ecological Modelling* 3: 309-313.

Fotheringham A.S., Brunsdon C., 2010. Local forms of spatial analysis. *Geographical Analysis* 31(4): 340-358.

Fotheringham A.S., Brunsdon C., Charlton M., 2002. Geographically weighted regression. Wiley, New York. 284 p.

Fotheringham A.S., Charlton M.E., Brunsdon C., 1998. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environmental and Planning A* 30(11): 1905-1927.

Gardner W., Mulvey E.P., Shaw E.C., 1995. Regression analyses of counts and rates: Poisson, over-dispersed Poisson, and negative binomial models. *Psychological Bulletin* 118(3): 392-404.

Gill A.M., Christian K.R., Moore P.H.R., 1987. Bush fire incidence, fire hazard and fuel reduction burning. *Australian Journal of Ecology* 12: 299-306.

Global News, 2017. Available at: https://globalnews.ca/news/3921710/b-c-year-in-review-2017-wildfires/.

Guo L., Ma Z., Zhang L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. *Canadian Journal of Forest Research* 38(9): 2526-2534.

Guo F., Innes L.J., Wang G., Ma X., Sun L., Hu H., Su Z., 2015. Historic distribution and driving factors of human-caused fires in the Chinese boreal forest between 1972 and 2005. *Journal of Plant Ecology* 8: 480-490.

Guo F., Su Z., Wang G., Sun L., Lin F., Liu A., 2016a. Wildfire ignition in the forests of southeast China: Identifying drivers and spatial distribution to predict wildfire likelihood. *Applied Geography* 66: 12-21.

Guo F., Selvalakshmi S., Lin F., Wang G., Wang W., Su Z., 2016b. Geospatial information on geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese boreal forest. *Canadian Journal of Forest Research* 46: 582-594.

Guo F., Wang G., Su Z., Liang H., Wang W., Lin F., Liu A., 2016c. What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire* 25: 505-519.

Guo, F., Su, Z., Wang, G., Sun, L., Tigabu, M., Yang, X., Hu, H., 2017. Understanding fire drivers and relative impacts in different Chinese forest ecosystems. *The Science of the Total Environment* 605: 411-425.

Haining R., Law J., Griffith D., 2009. Modeling small area counts in the presence of overdispersion and spatial autocorrelation. *Computational Statistics and Data Analysis* 53(8): 2923-2937.

Hu H.Q., 2005. Wildfire ecology and management. Second Ed. China Forestry Publishing House, Beijing.

Hu T., Zhou G., 2014. Drivers of lightning- and human-caused fire regimes in the Great Xing'an mountains. *Forest Ecology and Management* 329: 49-58.

Huang B., Zhang H., Sun Z., Zhou L., 2015. Forest fire danger factors and their division in Shandong based on GIS and RS. *Transactions of Chinese Journal of Ecology* 34(5): 1464-1472.

Justice C.O., Giglio L., Korontzi S., Owens J., Morisette J.T., Roy D., Descloitres J., Alleaume S., Petitcolin F., Kaufman Y., 2002. The MODIS fire products. *Remote Sensing of Environment* 83: 244-262.

Koutsias N., Martínez-Fernández J., Allgöwer B., 2010. Do factors causing wildfires vary in space? Evidence from geographically weighted regression. *GIScience and Remote Sensing* 47: 221-240.

Lee D., 2011. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology* 2(2): 79-89.

Liu Z., Yang J., Chang Y., Weisberg P.J., He H.S., 2012. Spatial patterns and drivers of fire occurrence and its future trend under climate change in a boreal forest of Northeast China. *Global Change Biology* 18: 2041-2056.

Maingi J.K., Henry M.C., 2007. Factors influencing wildfire occurrence and distribution in eastern Kentucky, USA. *International Journal of Wildland Fire* 16: 23-33.

Mandallaz D., Ye R., 1997. Prediction of forest fires with Poisson models. *Canadian Journal of Forest Research* 27: 1685-1694.

Masters A.M., 1990. Changes in fire frequency in Kootenay National Park, Canadian Rockies. *Canadian Journal of Botany* 68: 1763-1767.

McCoy V.M., Burn C.R., 2005. Potential alteration by climate change for forest-fire regime in the boreal forest of central Yukon Territory. *Arctic* 58: 276-285.

McCulloch C.E., Searle S.R., 2001. Generalized, linear and mixed models. John Wiley & Sons, Inc., New York, USA. pp. 1-184.

Miller J.D., Safford H.D., Crimmins M., Thode A.E., 2009. Evidence for increasing forest fire severity in the Sierra Nevada and Southern Cascade Mountains, California and Nevada, USA. *Ecosystems* 12: 16-32.

Miranda B.R., Sturtevant B.R., Stewart S.I., Hammer R.B., 2012. Spatial and temporal drivers of wildfire occurrence in the context of rural development in northern Wisconsin, USA. *International Journal of Wildland Fire* 21: 141-154.

Myers R.H., Montgomery D.C., Vining G.G., 2002. Generalized linear models. John Wiley & Sons, New York, USA. pp. 100-194.

Nakaya T., Fotheringham A.S., Brunsdon C., Charlton M., 2005. Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine* 24: 2695-2717.

National Interagency Fire Center, 2017. Available at: https://www.nifc.gov/fireInfo/nfn.htm.

Oliveira S., Oehler F., San-Miguel-Ayanz J., Camia A., Pereira J., 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and Random Forest. *Forest Ecology and Management* 275: 117-129.

Pastor E., Zarate L., Planas E., Arnaldos J., 2003. Mathematical models and calculation systems for the study of wildland fire behavior. *Progress in Energy and Combustion Science* 29: 139-153.

Potts J.M., Elith J. 2006. Comparing species abundance models. *Ecological Modeling* 199(2): 153-63.

Power J.H., Moser E.B., 1999. Linear model analysis of net catch data using the negative binomial distribution. *Canadian Journal of Fisheries and Aquatic Sciences* 56(2): 191-200.

Preisler H.K., Brillinger D.R., Burgan R.E., Benoit J.W., 2004. Probability based models for estimation of wildfire risk. *International Journal of Wildland Fire* 13(2): 133-142.

Randerson J.T., van der Werf G.R., Collatz G.J., Giglio L., Still C.J., Kasibhatla P., Miller J.B., White J.W.C., DeFries R.S., Kasischke E.S., 2005. Fire emissions for C3 and C4 vegetation and their influence on interannual variability of atmospheric CO_2 and $\delta^{13}CO_2$. *Global Biogeochemical Cycle* 19: GB2019.

Rodrigues M., de la Riva J., Fotheringham S., 2014. Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. *Applied Geography 48*: 52-63.

SAS Institute, Inc., 2013. STAT 9.4 Users' Manual. SAS Institute, Inc. Cary, USA.

Scholze M., Knorr W., Arnell N.W., Prentice I.C., 2006. A climate-change risk analysis for world ecosystems. *Proceedings of the National Academy of Science of the United States of America* 103: 13116-13120.

Scott J.H., Thompson M.P., Calkin D.E., 2013. A wildfire risk assessment framework for land and resource management. General Technical Report RMRS-GTR-315. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, p 83.

Sebastián-López A., Salvador-Civil R., Gonzalo-Jiménez J., San-Miguel Ayanz J., 2008. Integration of socio-economic and environmental variables for modelling long-term fire danger in Southern Europe. *European Journal of Forest Research* 127: 149-163.

Seyoum A., Ndlovu P., Zewotir T., 2016. Quasi-Poisson versus negative binomial regression models in identifying factors affecting initial CD4 cell count change due to antiretroviral therapy administered to HIV-positive adults in North–West Ethiopia (Amhara region). *AIDS Research and Therapy* 13: 36.

Spessa A., McBeth B., Prentice C., 2005. Relationships among fire frequency, rainfall and vegetation patterns in the wet-dry tropics of northern Australia: An analysis based on NOAA-AVHRR data. *Global Ecology and Biogeography* 14(5): 439-454.

Stocks B.J., Martell D.L., 2016. Forest fire management expenditures in Canada: 1970-2013. *Forestry Chronicle* 92(3): 298-306.

Syphard A.D., Radeloff V.C., Keely J.E., Hawbaker R.J., Clayton M.K., Stewart S.I., Hammer R.B., 2007. Human influence on California Fire Regimes. *Ecological Applications* 17: 1388-1402.

Syphard A.D., Radeloff V.C., Keuler N.S., Taylor R.S., Hawbaker T.J., Stewart S.I., Clayton M.K., 2008. Predicting spatial patterns of fire on a southern California landscape. *International Journal of Wildland Fire* 17: 602-613.

Terceiro M., 2003. The statistical properties of recreational catch rate data for some fish stocks off the northeast US coast. *Fishery Bulletin* 101: 653-672.

van Wilgen B.W., Biggs H.C., O'Regan S.P., Mare N., 2000. A fire history of the savanna ecosystems in the Kruger National Park, South Africa, between 1941 and 1996. *South Africa Journal of Science* 96(4): 167-178.

Vega Garcia C., Woodard P.M., Titus S.J., Adamowicz W.L., Lee B.S., 1995. A logit model for predicting the daily occurrence of human caused forest fires. *International Journal of Wildland Fire* 5(2): 101-111.

Ver Hoef J.M., Boveng P.L., 2007. Quasi-Poisson vs. negative binomial regression: how should we model over dispersed count data? *Ecology* 88(11): 2766-72.

Wotton B.M., Martell D.L., Logan K.A., 2003. Climate change and people-caused forest fire occurrence in Ontario. *Climatic Change* 60(3): 275-295.

Wotton B.M., Nock C.A., Flannigan M.D., 2010. Forest fire occurrence and climate change in Canada. *International Journal of Wildland Fire* 19(3): 253-271.

Wu W., Zhang L., 2013. Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in northeastern Puerto Rico. *Applied Geography* 37: 52-62.

Xiao Y., Ju H., Zhang X., Ji P., 2011. Relationship between fire-danger weather and forest fire in Qiannan area. *Scientia Silvae Sinicae* 47: 128-133 (in Chinese).

Xystrakis F., Koutsias N., 2013. Differences of fire activity and their underlying factors among vegetation formations in Greece. *iForest-Biogeosciences and Forestry* 6: 132-140.

Zhen Z., Shao L., Zhang L., 2018. Spatial hurdle Models for predicting the number of children with lead poisoning. *International Journal of Environmental Research and Public Health* 15(9): 1792.

Chapter III: Exploring Spatially Varying Relationships between Forest Fires and Environmental Factors at Different Quantile Levels

Abstract. In practice, forest fire managers have great interest in the impacts of driving factors on forest fire occurrence at various risk levels beyond an average fire risk. Using the forest fire occurrence data collected in Fujian province, P.R. China, we applied global quantile regression (QR) and geographically weighted quantile regression (GWQR) to investigate the spatially varying relationships between forest fires and environmental factors at the different quantiles (e.g., 50th, 75th, 90th, and 99th) of fire occurrence. Our results indicated that (1) at each quantile, the regression coefficients of both global QR and GWQR models were negative for elevation, slope, NDVI, and positive for the settlement density, national road density, and grass cover; (2) the low frequency of high fire occurrence events may dramatically affect the analyses and modeling on the relationships between fire occurrence and a specific environmental factor; (3) GWQR indicated that the relationships between forest fires and environmental factors significantly varied across the study area at different quantiles of forest fires; and (4) the GWQR models performed better in model fitting and prediction than the QR models at all quantiles. Therefore, the GWQR models provided a more complete view of forest fire distribution and highlighted the high risky locations of forest fires across the Fujian province, which would help the government agencies to make better decisions on where and what the fire management and prevention should focus on.

Keywords: forest fire count; risk assessment; quantile regression; geographically weighted quantile regression.

1. Introduction

Forest fires burn vegetation layers partially or completely and affect post-fire soil and vegetation processes such as soil erosion, debris flow, flooding, vegetation recovery, and changes in biodiversity (Scott *et al*. 2013). Forest fires can be extremely destructive, killing people, and destroying homes and other structures when they occur in wildland-urban interfaces. Although forest fires are inevitable, the destruction and damage from the hazards can be lowered through the principles of decision science and risk management (Calkin *et al*. 2014). Wybo *et al*. (1995) believed that the key point for prevention and firefighting was risk assessment. It can be done from different points of view and at different time scales, such as from historical data, by real time monitoring, and/or by forecasting (Blanchi *et al*. 2002). In addition to forest fire danger index systems (e.g., Bradshaw *et al*. 1984; Burgan *et al*. 1998; Lopez *et al*. 2002), statistical modeling has been applied and played an important role for the risk assessment of forest fires (Brillinger *et al*. 2003). Logistic regression was usually fitted to estimate the probability of forest fire ignition (Preisler *et al*. 2004; Preisler and Westerling 2007). Poisson regression was often utilized for estimating the number of forest fire occurrence (Mandallaz and Ye 1997; Wotten *et al*. 2010). In other cases, multiple linear regression was used (e.g., McKenzie *et al*. 2000; Syphard *et al*. 2007; Oliveira *et al*. 2012).

However, most published regression models are global in nature, assuming that the relationships between response and predictors are spatial stationary and / or homogenous. These assumptions may not be reasonable and appropriate for identifying the relationships between forest fire ignition / occurrence and influence factors across a large geographical region (Koutsias *et al*. 2010; Finney *et al*. 2011). On the other hand, geographically weighted regression (GWR) has become popular in different disciplines in recent decades (Fotheringham *et al*. 1998;

Foody 2003). Later, the framework of GWR has been extended to generalized linear models, such as geographically weighted logistic regression (GWLR) and geographically weighted Poisson regression (GWPR) (Nakaya *et al*. 2005). Previous studies in forest fire modeling usually applied the GWLR or/and GWPR method to explore the relationships between fire occurrence and regional variations of driving environmental factors (Koutsias *et al*. 2010; Martínez-Fernández *et al*. 2013; Rodrigues *et al*. 2014; Oliveira *et al*. 2014; Rodrigues *et al*. 2018).

Furthermore, most regression models, including global, GWR, and GWR generalized models, focus on the “mean or average” relationships between response variable and predictor variables so that they provide the prediction on the conditional mean (i.e., central behavior) of the response variable given the values of the predictors (Yu *et al*. 2003). In contrast, quantile regression (QR) provides the ability of exploring more complete and comprehensive picture of relationships between response variable and predictor variables (Koenker and Bassett 1978). It has been applied to various study fields such as ecology, investment, finance, economics, medicine, engineering, and etc. (e.g., Cade and Noon 2003; Zhang *et al*. 2005; Huang *et al*. 2017), as well as in wildfire studies (Mueller *et al*. 2014; Barros and Pereira 2014; Rijal 2018). Thus, QR can be particularly useful when people are interested in the relationships between forest fires and driving environmental factors at different risk levels and/or extremes of a fire response variable (e.g., probability of ignition or occurrence). Recently, Chen *et al*. (2012) developed geographically weighted quantile regression (GWQR) to integrate the GWR methodology with the traditional QR framework. This innovative approach provides a foundation for modeling spatially nonstationary relationships between variables at a range of conditional quantiles of the response variable distributions. Since then, to our best knowledge, there were very limited applications of GWQR in ecosystems and wildfire management.

In China, Fujian province ranks the highest forest coverage in the nation, but experiences high annual forest fire incidences, with nearly 15,000 forest fires occurring from 2000 to 2010. Although the fire prevention efforts have reduced the number of annual forest fires in recent years, the total area of forest burned has increased in the Fijian province (Guo *et al*. 2018). However, the analysis of influencing factors on forest fires is a relatively new and developing study field in China, which has been mainly focused on the boreal forests in northern China (Wu *et al*. 2014; Guo *et al*. 2015; Guo *et al*. 2016b). Up to date, limited modeling approaches have been utilized to develop statistical models for the risk assessment and prevention of forest fires in China, including logistic regression (Guo *et al*. 2015), multiple linear regression (Liu *et al*. 2012), and random forests (Wu *et al*. 2014; Guo *et al*. 2016c; Guo *et al*. 2017), which results in less informative fire management plan, especially in the subtropical regions like the Fujian province.

The objective of this study was to apply global quantile regression (QR) and geographically weighted quantile regression (GWQR) to model the spatially varying forest fires against the environment factors of infrastructure, topography, and meteorology at different quantiles, rather than the conditional mean / average, of the forest fire occurrence. Hopefully, these quantile models enable us to explore the full distribution of forest fires and identify “high fire occurrence” locations or areas across the Fujian province, China.

2. Data

2.1 Study area

Fujian province is located in a sub-tropical region of China with a total land area of 124,000 km^2 (Figure 3.1). It ranks the highest forest coverage in the nation (about 66% of Fujian province is covered by forests and vegetation), but experiences high annual forest fire incidences, with nearly 15,000 forest fires occurring from 2000 to 2010 (Guo *et al*. 2018). The dominant tree species in Fujian include Massoniana (*Pinus massoniana* Lamb.), Chinese fir (*Cunninghamia lanceolate* (Lamb.) Hook), Casuarina (*Casuarina equisetifolia* L.), and Pubescens (*Phyllostachys heterocycle* (Carr.) Mitford cv. *Pubescens*). The climate is warm and humid with an average annual rainfall of 1400 – 2000 mm and average temperature of 17 – 21 °C. Forest fire season is typically spanning from September to April (Guo *et al.* 2016a).

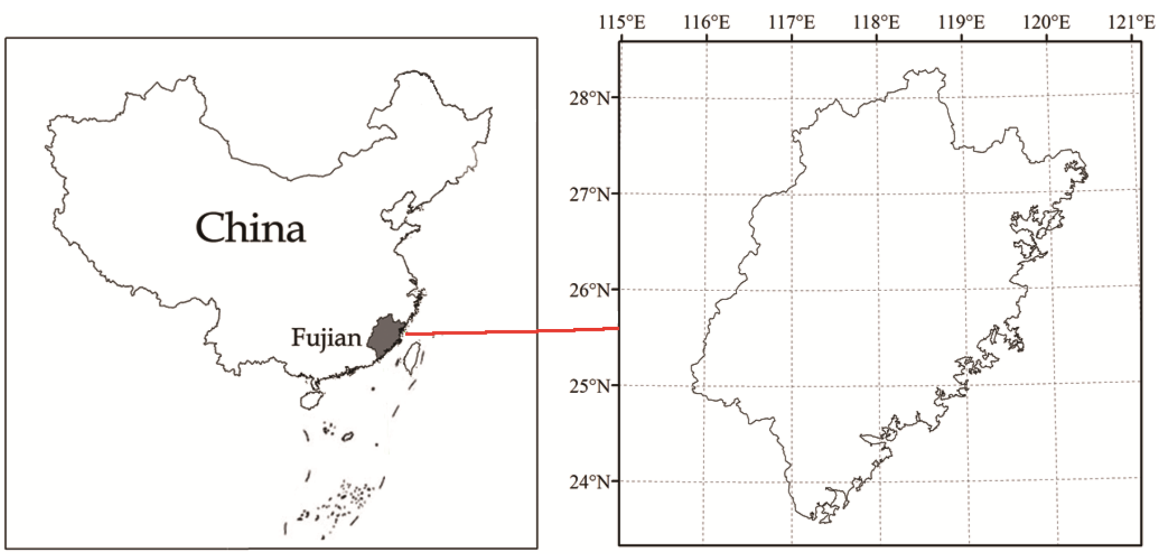


Figure 3.1 Location map of Fujian province, P.R. China.

2.2 Data collection

2.2.1 Fire data (response variable)

In this study, we used MODIS hotspots product (MOD14A1) which has been considered as a reliable and suitable source for monitoring forest fires to analyze the relationship between forest fire occurrence and environmental factors in Fujian, China (Guo *et al*. 2016 a,b; Su *et al*. 2019). The time span of the study is 16 years (2001-2016). Since MOD14A1cannot distinguish forest fires from non-forest fires that occur in cities/towns, construction sites, agricultural lands, and other areas, we further processed the fire data by: (1) removing the fire points in cities/towns, construction sites, and farmland based on a 1 km resolution land-use map; and (2) extracting fire points based on the time of fire occurrence within the fire season of the study area (September 15 to April 30 of the following year). The whole study area was divided into 4 × 4 km grids (a total of 7433 grids) using ArcGIS 10.2 (ESRI 2010) and the total number of forest fire occurrences in each grid were calculated as the response variable in the model fitting. The quantiles of forest fire occurrence were calculated and summarized in Table 3.1. The frequency and spatial distributions of the forest fire occurrence are illustrated in Figure 3.2.

Table 3.1 Distribution summary of the forest fire occurrence.

$\boldsymbol{\tau}$	**0.25**	**0.50**	**0.60**	**0.75**	**0.80**	**0.90**	**0.95**	**0.99**	**1.0**
$\xi(\tau)$	1	3	3.2	6	7	10	14	27	153

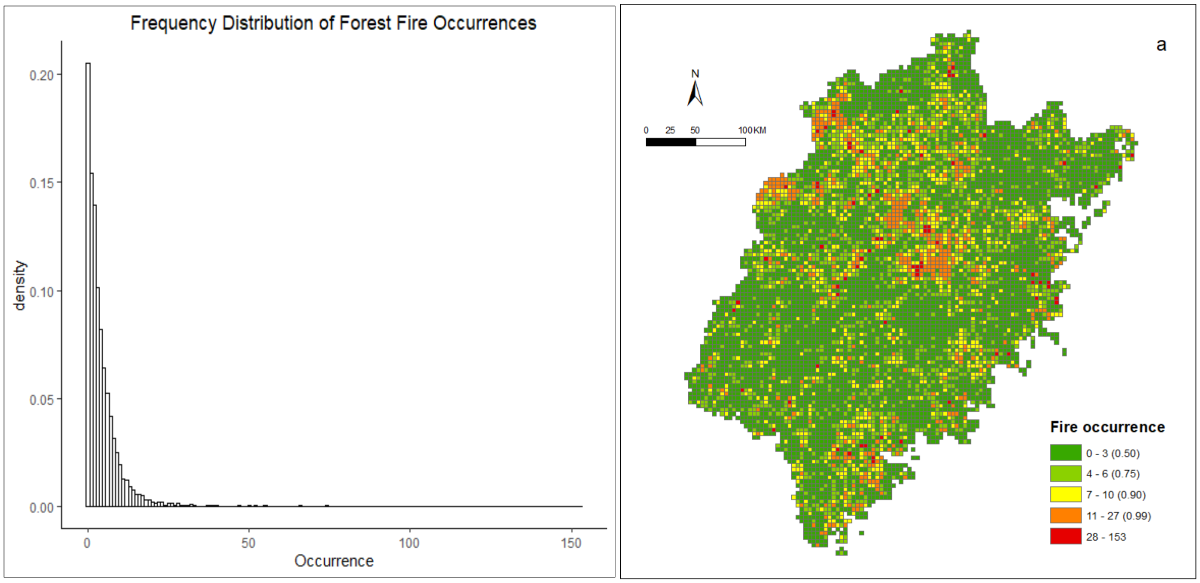


Figure 3.2 Frequency and spatial distributions of the forest fire occurrence.

2.2.2 Potential driving factors (predictor variables)

A total of 18 predictor or explanatory variables were collected and grouped into four categories, including topographical, meteorological, human related, and vegetation and land use predictors. The specific variable collection processes are as follows:

Topographic variables

Topographic variables included elevation (km), slope (degree), and aspect index. High resolution (25 m) Digital Elevation Model (DEM) data were collected from the National Administration of Surveying, Mapping and Geoinformation of China (http://www.gscloud.cn/sources). The slope and slope direction were derived from DEM, and aspect was then converted into an aspect index using the following formula: Aspect Index = $\cos(\theta \times \pi / 180)$, where θ is the degree of slope generated in ArcGIS ranging from 0 – 360° so that the aspect index ranges from -1 to 1 (Guo *et al.* 2017). The average elevation, slope, and aspect index of each grid were then extracted using ArcGIS 10.2.

Meteorological variables

Meteorological variables included precipitation (mm/day), temperature (°C), and relative humidity (%), which were obtained from the platform of National Earth System Science Data Center (http://www.geodata.cn), an important component of National Science and Technology Infrastructure. The climatic variables are interpolated from ANUSPLIN, a software package developed by Hutchinson (2004) based on the thin-plate smoothing method to generate hydrometeorological maps. ANUSPLIN includes a linear covariate to represent the elevation dependent meteorological factors, and it outperformed in climate interpolation (Zhang *et al*. 2010) and long period climatic data (McKenney *et al*. 2006). Raster calculator in ArcGIS 10.2 was used to calculate the annual average of each meteorological variable for each grid from year 2001 to 2016. Precipitation and temperature impact the occurrence of forest fire by limiting the fuel moisture content. Therefore, they are reasonable and effective alternative fuel factors when other fuel factor is not available. In addition, the annual meteorological factor is a traditional and better indicator to measure the influence of climate change on forest fires, compared to the average meteorological data during the period of forest fire (Scholze *et al*. 2006; McCoy and Burn 2005; Xystrakis *et al.* 2013).

Human factors

Human factors included the socioeconomic variables (per capita GDP and population density) and infrastructural variables. The GDP and population data were obtained from Resource and Environmental Data Cloud Platform (http://www.resdc.cn/Default.aspx) and the data resolution was 1 km. The infrastructural variables included road density (km/km^2, ratio of road length to the grid area) and water density. A 1:250,000 vector map of infrastructure was provided by the National Geomatics Center of China (http://www.ngcc.cn/). We classified the road into national, provincial, and local road. Their buffer areas were built based on 50 m, 25 m, and 10 m, respectively, by using the tool of neighborhood analysis in ArcGIS 10.2. The ratio of the road area was calculated in each grid (Hoyo *et al*. 2011). All the selected human factors for each grid from year 2001 to 2016 were calculated using ArcGIS 10.2.

Vegetation coverage and land use factor

The Normalized Difference Vegetation Index (NDVI) was used to reflect the vegetation coverage of the study area. The NDVI data were derived from the MODIS NDVI product with a spatial resolution of 500 m provided by the Geospatial Data Cloud (http://www.gscloud.cn/). The land use data (1 km resolution) were obtained from the Resource and Environmental Data Cloud Platform (http://www.resdc.cn/Default.aspx). It provides the spatial distribution of vegetation types by digitizing the collections of vegetation type in China on the scale 1 : 1 million. Forest, including subtropical evergreen broad-leave forest type and mixed conifer and broad-leave forest type, covers about 64.95% of the total area. Vegetation type of shrub contains the subtropical evergreen broad-leave shrub, tropical evergreen broad-leaved shrub, and deciduous and broad-leaved shrub, taking about 20.40% of the study region. The cultivated land, fruit forest, and non-timber product forest are categorized into the cropland, covering about 12.6% of the total area. Grass (subtropical grass and tropical grass) only covers 1.69% of the total area. About 0.36% of the total area is the development land. According to the survey, regional forest coverage proportion is 66.80% and the agriculture land is about 11%, indicating the land classification of the raster data is close to reality. ArcGIS10.2 was used to calculate the proportion of each land cover in each grid.

2.2.3 Multicollinearity analysis among explanatory variables

We used the variance inflation factor (VIF) to detect the multicollinearity among variables before fitting the regression models. In general, a VIF above 10 indicates high correlations between explanatory variables (Guo *et al*. 2017). In this study, the socioeconomic variable of GDP was removed because its VIF was 18.58, while other 17 explanatory variables were used to fit both global and GWR quantile models. The descriptive statistics of the response variable (i.e., forest fire occurrence) and 17 predictor variables were listed in Table 3.2. The spatial distributions of 17 predictor variables across the Fujian province are shown in Figure 3.3.

Table 3.2 Descriptive statistics of the response and predictor variables.

Variable	Mean	Median	Std Dev	Minimum	Maximum
Fire Occurrence	4.2500	3.000	6.4271	0.0000	153.000
Elevation (km)	0.4885	0.4688	0.2714	-0.0064	1.7577
Slope (degree)	19.78	20.72	5.77	0.105	37.68
Aspect Index	-0.0073	-0.0116	0.1186	-0.5728	0.5523
Precipitation (mm/day)	1670	1682	158.1292	1247	2042
Temperature (°C)	18.3	18.0	1.1564	15.27	21.12
Humidity (%)	76.07	76.00	1.4118	73.00	79.20
River Density (%)	0.5969	0.5584	0.1962	0.0015	1.6537
Settlement Density (%)	0.4523	0.0000	3.7411	0.0000	95.4832
National Road Density (%)	0.2006	0.0000	0.7451	0.0000	6.5424
Provincial Road Density (%)	0.2176	0.0000	0.5220	0.0000	5.3110
Local Road Density (%)	1.052	1.051	0.5848	0.0000	4.873
Population (people/km^2)	283.9	133.5	535.7	56.8	17899.5
NDVI	0.7742	0.7910	0.0662	0.3287	0.8729
Forest Cover (%)	51.752	55.374	39.487	0.000	100.00
Shrub Cover (%)	18.608	0.000	28.160	0.000	99.99
Grass Cover (%)	1.409	0.000	9.583	0.000	99.99
Crop Cover (%)	10.161	0.000	22.499	0.000	99.99

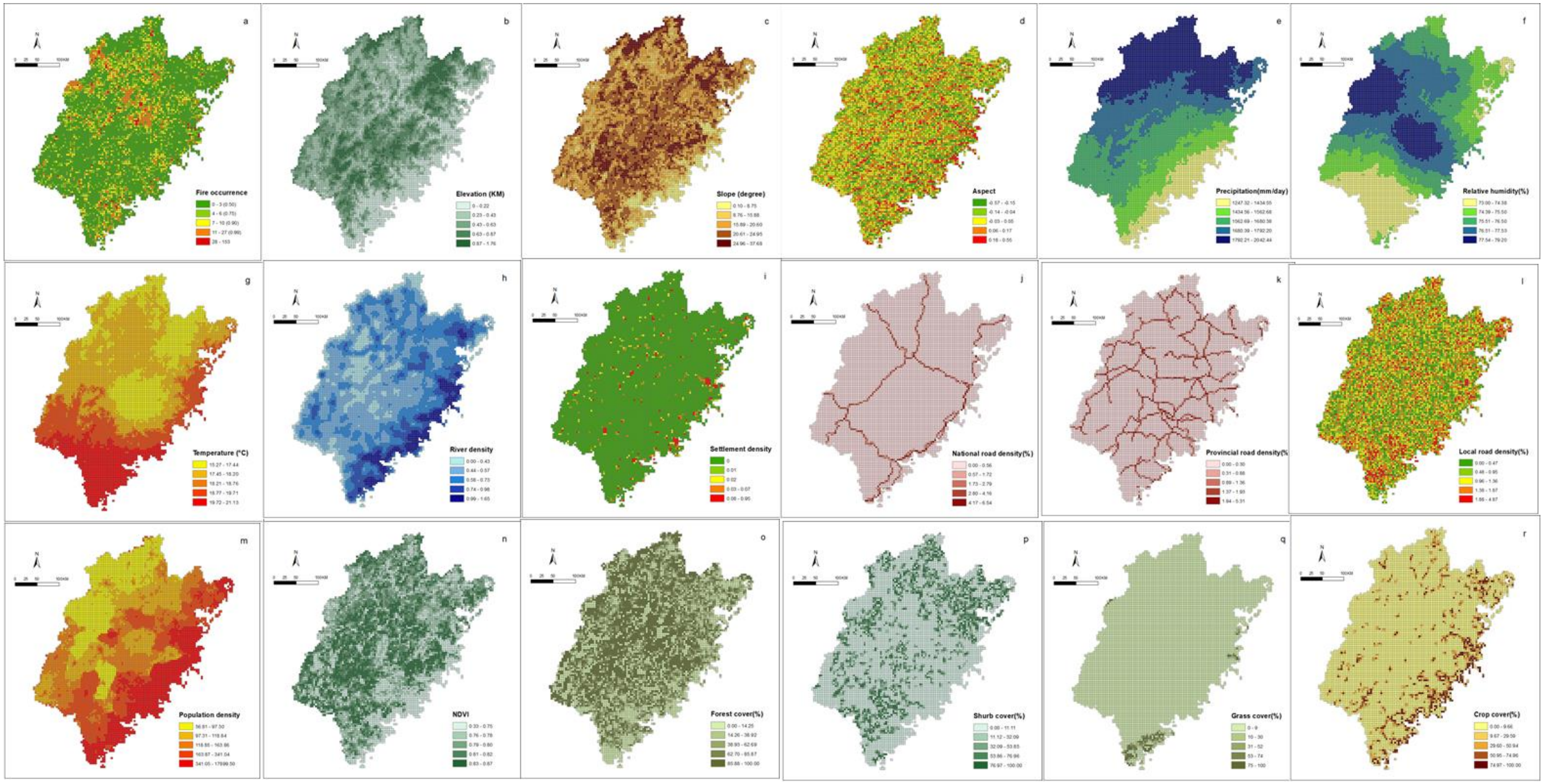


Figure 3.3 Spatial distributions of (a) Forest fire counts classified by quantiles, (b) Elevation, (c) Slope, (d) Aspect index, (e) Precipitation, (f) Relative humidity, (g) Temperature, (h) River density, (i) Settlement density, (j) National road density, (k) Provincial road density, (l) Local road density, (m) Population density, (n) NDVI, (o) Forest cover, (p) Shrub cover, (q) Grass cover, and (r) Crop cover.

3. Method

3.1 Theoretical background

3.1.1 Quantile regression (QR)

For a random response variable Y with a cumulative distribution function (***cdf***) $F(y) = \Pr(Y \le y)$, the $\boldsymbol{\tau}$th quantile of Y is defined as the inverse of the ***cdf*** at $\boldsymbol{\tau}$, that is the value of Y such that $F(Y) = P(Y \le \xi) = \tau$, where $0 < \boldsymbol{\tau} < 1$. Thus, the proportion of the population with the response variable below $\xi(\tau)$ is $\boldsymbol{\tau}$. The inverse function $Q(\tau) = F^{-1}(\tau) = \inf(y : F(Y) \ge \tau)$ is called the quantile function of F(Y). The general $\boldsymbol{\tau}$th sample quantile $\xi(\tau)$, which is the analogue of $Q(\tau)$, can be obtained by minimizing:

$$\xi(\tau) = \min \sum_{i=1}^{n} \rho_\tau\left(Y_i - \xi(\tau)\right) = \min\left[(1-\tau)\sum_{Y_i < \xi}\left(Y_i - \xi(\tau)\right) + \tau \sum_{Y_i \ge \xi}\left(Y_i - \xi(\tau)\right)\right] \qquad [1]$$

where the loss function ρ_τ assigns a weight of $\boldsymbol{\tau}$ to positive residuals $Y_i - \xi(\tau)$ and a weight of $(1 - \tau)$ to negative residuals (Koenker 2005).

QR is designed to model the effects of ***p*** predictor variables (**X)** on the conditional percentiles or quantiles of a response variable, such that $Q(\tau \mid X) = X\beta(\tau) + \varepsilon(\tau)$, where $0 < \boldsymbol{\tau} < 1$. However, the model error term $\boldsymbol{\varepsilon(\tau)}$ is unspecified and is only assumed that $\boldsymbol{\varepsilon(\tau)}$ satisfies the quantile restriction $Q(\varepsilon(\tau) \mid X) = 0$ (Koenker and Bassett 1978). The QR coefficients can be obtained by solving for any quantile $0 < \boldsymbol{\tau} < 1$:

$$\hat{\beta}(\tau) = \min \sum_{i=1}^{n} \rho_\tau\left(Y - X\beta(\tau)\right) \qquad [2]$$

where ρ_τ is a V-shaped piecewise loss function (Koenker 2005). For the case of $\boldsymbol{\tau} = 0.5$, the QR is the median regression, also known as $\mathbf{L_1}$ regression estimator.

3.1.2 Geographically weighted quantile regression (GWQR)

To investigate the spatial heterogeneity of a regression relationship, the data must be collected with the location coordinates (u_i, v_i) for each observation ***i*** (***i*** = 1, 2,…, *n*). The local information leads to estimate the localized regression coefficients of the relationship of interest. Chen *et al*. (2012) expressed the GWQR as follows:

$$Y_i(\tau) = X_i\beta(\tau)(u_i, v_i) + \varepsilon_i(\tau) = \beta_0(\tau)(u_i, v_i) + \sum_{k=1}^{p}\beta_k(\tau)(u_i, v_i)X_{ki} + \varepsilon_i(\tau) \qquad [3]$$

where Y_i is the response variable, X_k is a set of ***p*** predictor variables (***k*** = 1, 2,…, *p*), ε_i is the error term of the conditional ***τ***th quantile function, and $\beta_0(\tau)(u_i, v_i), \beta_1(\tau)(u_i, v_i), \cdots, \beta_p(\tau)(u_i, v_i)$ are the local QR coefficients for the ***τ***th quantile at the ***i***th location.

For a given regression point (u_0, v_0), the solution of the GWQR coefficients for the ***τ***th quantile in Eq. [3] can be obtained by minimizing the geographically weighted loss function using the data within the kernel window:

$$\sum_{i=1}^{n}\rho_\tau\left(Y_i(\tau) - \beta_0(\tau)(u_0, v_0) - \sum_{k=1}^{p}X_{ik}\beta_k(\tau)(u_0, v_0)\right)\cdot W_0 \qquad [4]$$

where W_0 is the spatial weights defined by a kernel function K(***d_{0i}*** / ***h***), where ***h*** is the bandwidth and ***d_{0i}*** is the distance between each neighboring location ***i*** and the regression point (u_0, v_0). Note: there is no explicit form available for the solution of the model coefficient vector in Eq. [4]. Instead, it can be equivalently formulated as a linear programming optimization problem (Chen and Wei 2005; Koenker 2005) as follows: let $(u_0, v_0) = (u_i, v_i)$ (*i* = 1, 2,…, *n*), the estimator $\hat{\beta}_k(\tau)(u_i, v_i)$ (***k*** = 0, 1, 2,…, ***p***) for each GWQR coefficient can be obtained so that the corresponding ***τ***th quantile is estimated by

$$\hat{Q}_{\tau}(X_i, u_i, v_i) = X_i\hat{\beta}(\tau)(u_i, v_i) = \hat{\beta}_0(\tau)(u_i, v_i) + \sum_{k=1}^{p} X_{ik}\hat{\beta}_k(\tau)(u_i, v_i) \quad [5]$$

where $\hat{\beta}(\tau)(u_i, v_i)$ is the vector of regression coefficient estimates and X_i denotes the vector of observed predictor variables at the ***i***th location (u_i, v_i). More details on the GWQR coefficient estimates, standard errors of regression coefficients, kernel function, bandwidth selection, and the assessment of spatial non-stationarity were available in Chen *et al*. (2012).

3.2 Regression model

We chose the following linear models for both global QR and GWQR to explore the relationships between forest fire occurrence (Y_i) and predictor variables at the four quantiles of Y_i (τ = 0.50, 0.75, 0.90, and 0.99), respectively:

$$Y_i(\tau) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{16} X_{16i} + \beta_{17} X_{17i} + \varepsilon_i \quad [6]$$

$$Y_i(\tau) = \beta_0(u_i,v_i) + \beta_1(u_i,v_i)X_{1i} + \beta_2(u_i,v_i)X_{2i} + \ldots + \beta_{17}(u_i,v_i)X_{17i} + \varepsilon_i \quad [7]$$

The SAS procedure PROC QUANTREG was used to fit the global QR models (Eq. [6]) (SAS Institute, Inc. 2013), and the SAS macro provided by Chen *et al*. (2012) was used to fit the GWQR models (Eq. [7]).

3.3 Bandwidth selection for GWQR

In this study, we used Akaike Information Criterion (AIC) to decide the optimal bandwidth and related kernel function for estimating each regression coefficient for each geographic location *i* and each predictor variable (Fotheringham *et al*. 2002; Guo *et al*. 2008). Variogram model (Bailey and Gatrell 1995) was fitted through the variogram values in order to find the optimal bandwidth and kernel function. We tried the residuals of global quantile models at the four quantiles for variogram respectively. All optimal kernel functions and estimated bandwidth were chosen at the smallest AIC.

3.4 Assessment of spatial autocorrelation and nonstationary

Existence of spatial autocorrelation and heterogeneity can be evaluated from the model residuals of the four global quantile regression models by Moran's Index (Moran 1950). The positive Moran's I values indicated that a "high fire occurrence number (HON)" pixel was neighboring with the HON pixels, while a "low fire occurrence number (LON)" pixel was neighboring with the LON pixels across the study area.

To evaluate the spatial variation in regression coefficients of GWQR, we followed the approach in Chen *et al*. (2012). At a specified quantile, the interquartile ranges (IQR) of the local model coefficients computed by GWQR were compared with the corresponding standard errors of the global QR model coefficients. When IQR was twice as large as the standard error, it would indicate that spatial non-stationarity existed in the relationship between response variable and a specific predictor variable.

3.5 Model evaluation

The pinball loss function was used to evaluate the prediction accuracy of the quantile regression models. Different from the classic regression models, in which the goal is to have the model prediction as close as possible to the observed values of response variable, quantile regression is designed to estimate the conditional quantiles by minimizing asymmetrically weighted errors, which is called pinball loss. It returns a value interpreted as the accuracy of a quantile regression model. The lower the pinball loss is, the more accurate the quantile model is (Yu *et al*. 2018). The pinball loss function in quantile regression can be expressed as:

$$L_\tau = (Y - Z)\,\tau \qquad \text{if} \quad Y \geq Z$$

$$L_\tau = (Z - Y)(1 - \tau) \qquad \text{if} \quad Y < Z \qquad [8]$$

where Y represents the observed quantile value and Z is the predicted quantile value at the target quantile τ.

4. Results

4.1 Relationships between forest fires and environmental factors based on global QR

The global QR models were fitted with the seventeen predictor variables at τ = 0.50, 0.75, 0.90, and 0.99 quantiles, respectively. In terms of statistical testing on the regression coefficients, the predictor variables of elevation, slope, and NDVI were statistically significant at all four quantiles, while local road density, forest cover, and shrub cover was statistically non-significant at the significance level α = 0.05 (Table 3.3). The significance of other predictors varied depending on a particular quantile. For example, the predictors of precipitation, humidity, and population density were significant at τ = 0.50, 0.75 and 0.95 quantiles, but not significant at τ = 0.99 quantile. Other predictor variables were only important at a particular quantile (Table 3.3).

The regression coefficients of three topographical variables (elevation, slope, and aspect index), NDVI, and crop cover were negative at all quantiles, indicating that the forest fire occurrence reduced with higher elevation, steeper terrain, denser vegetation and / or crop cover. In contrast, precipitation, settlement density, and national road density were positively related to the forest fire occurrence, implying that heavier rainfall, denser settlement and larger national road occupancy may cause higher chance of forest fire occurrence (Table 3.3). Some variables such as precipitation, settlement density, humidity, and population density were significant at τ = 0.50, 0.75, and 0.90 quantiles, but not at τ = 0.99 quantiles. In particular, settlement density was significant at low quantiles (0.50, 0.75), indicating settlement density had more contribution in affecting the LON pixels (i.e., less forest fire occurrences). Inversely, grass cover was not significant at τ = 0.50 quantile, but strongly related to the HON pixels (Table 3.3). The spatial maps of the coefficients of each predictor variable at the four quantiles confirmed that the relationships between fire occurrence and some predictors were fluctuated as the quantile increased (Figure 3.4).

Table 3.3 Global quantile regression (QR) estimates for the 0.50, 0.75, 0.90, and 0.99 quantiles.

Parameters	$\tau = 0.50$			$\tau = 0.75$			$\tau = 0.90$			$\tau = 0.99$		
	Estimate	**Standard Error**	***p*-value**	**Estimate**	**Standard Error**	***p*-value**	**Estimate**	**Standard Error**	***p*-value**	**Estimate**	**Standard Error**	***p*-value**
Intercept	-23.972	5.6425	<.0001	-23.294	10.2676	0.0233	-73.928	16.6996	<.0001	107.340	92.1423	0.1605
Elevation	-3.7732	0.1758	<.0001	-6.8132	0.3157	<.0001	-8.9991	0.4444	<.0001	-18.6104	2.5360	<.0001
Slope	-0.0504	0.0144	<.0001	-0.0478	0.0236	0.0434	-0.196	0.0374	<.0001	-0.5613	0.1724	0.0065
Aspect Index	-1.1098	0.3368	0.0010	-1.3137	0.6183	0.0336	-1.7822	0.9224	0.0534	-2.8749	4.0387	0.4766
Precipitation	0.0027	0.0004	<.0001	0.0030	0.0007	<.0001	0.0066	0.0011	<.0001	0.0084	0.0052	0.1050
Humidity	0.331	0.0521	<.0001	0.45	0.091	<.0001	1.0062	0.1284	<.0001	-0.2806	0.7505	0.7085
Temperature	0.1379	0.0777	0.0759	0.0512	0.355	0.7057	0.6082	0.1933	0.0017	-1.2191	1.0559	0.2483
River Density	0.5599	0.3281	0.0880	0.9599	0.5352	0.0729	2.4154	0.9083	0.0079	-1.7924	4.0387	0.6880
Settlement Density	0.1196	0.0258	<.0001	0.3853	0.1236	0.0018	0.6683	0.4699	0.1500	0.7496	5.4149	0.8899
National Road Density	0.3034	0.0998	0.0024	0.5114	0.2168	0.0183	0.7225	0.3769	0.0553	6.8423	3.1275	0.0287
Provincial Road Density	-0.0043	0.1001	0.9657	-0.179	0.1721	0.4642	0.2404	0.5170	0.6420	14.9971	4.4116	0.0007
Local Road Density	-0.0202	0.0768	0.7927	-0.0997	0.1517	0.5111	-0.1298	0.2069	0.5304	-0.2647	0.7839	0.7356
Population	-0.0014	0.0002	<.0001	-0.002	0.0003	<.0001	-0.0028	0.0005	<.0001	-0.005	0.0041	0.2223
NDVI	-3.1728	1.3755	0.0211	-9.1403	2.5307	0.0003	-9.7537	4.2174	0.0208	-42.4522	11.8378	0.0003
Forest Cover	-0.1443	0.1315	0.2727	0.0906	0.2286	0.6920	-0.0134	0.3952	0.9730	-0.6700	1.2367	0.5880
Shrub Cover	0.1240	0.1679	0.4600	0.4191	0.2888	0.1468	-0.2387	0.4550	0.5998	-1.2901	1.9927	0.5174
Grass Cover	1.4590	1.0544	0.1665	2.9718	1.3932	0.0329	4.0933	1.8053	0.0234	7.0644	4.2930	0.0999
Crop Cover	-1.2928	0.2750	<.0001	-2.5684	0.4361	<.0001	-1.8676	1.0289	0.0695	-6.7250	3.3823	0.0468

Note: The quantile model coefficient indicates how much the response variable changes at a quantile τ for one unit change in a predictor variable. For example, at a moderate risk level ($\tau = 0.50$), the predicted fire occurrence will be decreased by 3.7732 when the elevation rises one kilometer, while at a high risk level ($\tau = 0.99$), the predicted forest fire occurrence will be decreased by 18.61 when the elevation rises one kilometer.

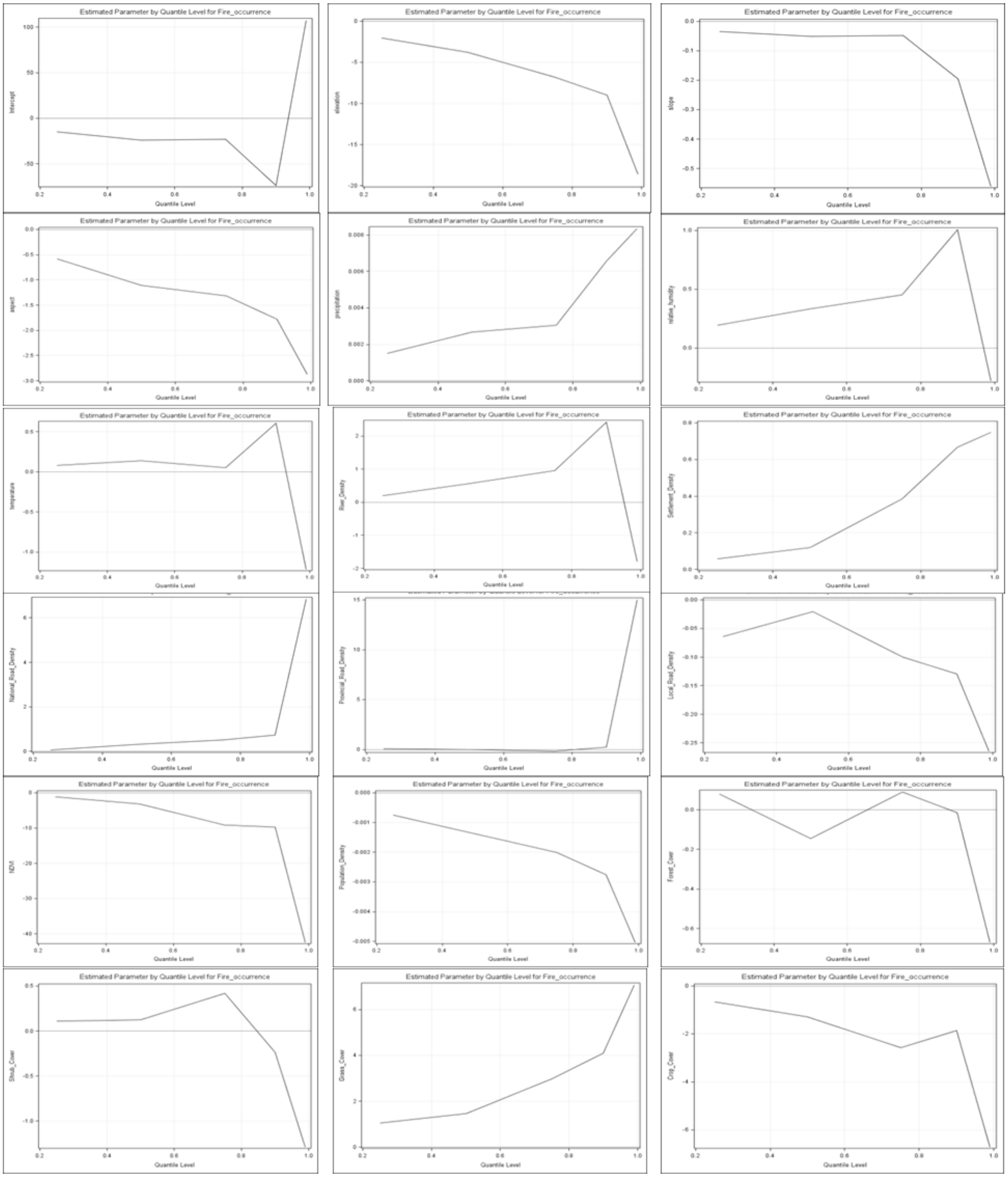


Figure 3.4 Model coefficient estimates of the global quantile regression models at different quantiles.

4.2 Spatial autocorrelation and nonstationary analysis

Table 3.4 revealed that the spatial autocorrelations were statistically significant in the study area. The Moran's I became greater in the model residuals for larger quantiles (e.g., τ = 0.90 and 0.99 quantiles), implying that the HON pixels (i.e., more forest fire occurrences) were more spatially clustered. We performed the variogram analyses on the residuals of global quantile models at the four quantiles. Specifically, the kernel functions and estimated bandwidths at τ = 0.50, 0.75, and 0.90 quantiles are similar, but dissimilar at τ = 0.99 quantile (Table 3.4).

In addition, our results indicated that most IQRs of the localized model coefficients were at least twice the standard errors of the global model coefficients, except for slope, aspect index, river density, provincial road density, local road density, and shrub cover at $\boldsymbol{\tau}$ = 0.99 quantile, suggesting that the relationships between forest fire occurrences and some environmental factors indeed varied across the Fujian province (Table 3.5).

Table 3.4 Moran's Index and variogram estimations for the residuals of global quantile regression models.

Statistics	Residuals $\tau = 0.50$	Residuals $\tau = 0.75$	Residuals $\tau = 0.90$	Residuals $\tau = 0.99$
Moran's Index	0.0165	0.0145	0.0210	0.0263
Z-score	64.21	57.81	83.40	104.2
p-value	<0.0001	<0.0001	<0.0001	<0.0001
Kernel	Spherical	Spherical	Spherical	Exponential
Bandwidth (m)	136,361	134,707	137,115	300,081

Table 3.5 Summary of model parameter estimation of geographically weighted quantile regression models.

Statistics	τ	$\beta_{Intercept}$	$\beta_{Elevation}$	β_{Slope}	$\beta_{Aspect\ index}$	$\beta_{Precipitation}$	$\beta_{Relative\ Humidity}$	$\beta_{Temperature}$	$\beta_{River\ density}$	$\beta_{Settlement\ density}$	$\beta_{National\ road\ density}$	$\beta_{Provincial\ road\ density}$	$\beta_{Local\ road\ density}$	$\beta_{Population}$	β_{NDVI}	$\beta_{Forest\ cover}$	$\beta_{Shrub\ cover}$	$\beta_{Grass\ cover}$	$\beta_{Crop\ cover}$
Mean	0.50	-13.51	-2.779	-0.0657	-0.7158	-0.0036	0.3271	0.3720	0.5404	0.1348	0.3971	-0.0336	0.021	-0.0008	-8.024	-0.1068	0.4426	5.747	-0.8276
	0.75	-19.52	-5.838	-0.0405	-1.5363	-0.0014	0.5249	0.2822	1.1442	0.7524	0.7609	-0.1928	-0.1732	0.0001	-17.594	0.2048	0.0573	6.5613	-1.3778
	0.90	-4.041	-7.698	-0.1133	-2.4093	-0.0033	0.5999	0.0088	1.372	1.828	1.9107	-0.0644	-0.3566	0.0026	-25.087	-0.5396	-0.7937	11.105	-2.0121
	0.99	195.9	-6.796	-0.379	1.1198	-0.0467	-0.04726	0.3227	4.139	3.038	3.6985	8.4171	-0.5973	-0.0012	-112.66	-1.4494	-0.4667	1.2935	5.909
Median	0.50	-12.75	-2.669	-0.0538	-0.8845	-0.0038	0.4358	0.2944	0.4364	0.0979	0.3708	-0.0826	0.0193	-0.0006	-8.355	-0.0856	0.1812	5.643	-0.6782
	0.75	-26.57	-5.519	-0.0202	-1.1639	-0.0024	0.6193	0.3839	0.7088	0.5125	0.7163	-0.2963	-0.2427	-0.0004	-16.44	0.1800	0.0754	5.2588	-1.0895
	0.90	-28.04	-7.648	-0.0952	-2.1791	-0.0045	0.8239	0.2206	1.306	1.3832	1.7431	-0.0061	-0.4442	0.0024	-26.282	-0.278	-0.5999	6.3304	-1.8226
	0.99	193.2	-6.187	-0.3973	0.1138	-0.0517	-0.2925	0.4252	5.388	2.795	3.4511	8.4302	-0.9466	-0.0018	-112.49	-0.6857	-0.5331	-0.6775	4.428
Minimum	0.50	-202.6	-9.649	-0.4293	-3.6184	-0.0206	-2.0287	-1.814	-2.4516	-0.1673	-0.7625	-1.1016	-1.0113	-0.0079	-43.849	-1.3544	-3.7099	-5.665	-4.5105
	0.75	-243.3	-16.85	-0.6173	-9.8455	-0.0308	-3.1822	-3.8673	-4.5859	-0.2564	-0.9994	-1.5665	-1.7811	-0.0126	-60.886	-2.3739	-2.2614	-30.225	-8.3854
	0.90	-422.5	-21.593	-1.0046	-10.737	-0.0436	-6.1908	-6.5644	-6.098	-3.0567	-1.0502	-3.2589	-2.7742	-0.014	-84.492	-6.4118	-8.3341	-67.269	-18.998
	0.99	33.62	-18.22	-0.6931	-7.8383	-0.1092	-4.4063	-12.085	-10.147	-1.457	-2.3926	-0.0584	-2.371	-0.0082	-277.92	-7.4125	-9.7132	-20.874	-12.486
Maximum	0.50	190.3	3.641	0.3187	2.6024	0.0111	2.1126	2.5463	4.4767	0.78	1.5756	1.2759	1.0538	0.0052	23.70	0.9167	3.1243	18.263	1.2993
	0.75	309.1	0.24	0.6869	5.2024	0.0235	2.7727	3.1958	1.49	0.0457	0.002	0.0482	0.0157	0.0099	34.208	2.4918	3.005	35.644	3.2066
	0.90	554.2	4.833	0.9434	4.9056	0.0314	4.9197	4.0218	17.215	6.8099	6.5476	3.19	2.3727	0.0155	105.20	1.7726	3.5661	60.654	8.4625
	0.99	798.2	6.461	0.0113	13.722	0.0159	3.6721	13.5014	14.831	7.826	12.0922	20.6273	2.4694	0.0088	61.09	1.9779	6.794	25.811	45.712
IQR	0.50	62.33	2.745	0.1391	1.2011	0.0092	0.8602	0.6814	2.1393	0.1138	0.2684	0.3746	0.2369	0.0026	11.909	0.4798	1.1189	5.9112	1.3949
	0.75	124.2	4.0153	0.1697	3.1509	0.0107	1.2015	1.4599	4.3478	0.7968	0.8263	0.7699	0.5746	0.0058	26.168	1.1741	1.5377	13.502	3.1648
	0.90	200.4	5.3934	0.3226	4.6259	0.0209	2.4057	2.4312	6.2844	1.8279	2.2765	1.6812	0.9823	0.0093	50.374	1.6239	2.1605	26.016	4.2225
	0.99	552.7	8.3711	0.2058	5.638	0.0404	3.1386	10.725	5.6906	3.5416	6.038	5.3343	1.2454	0.0044	122.34	3.5506	3.8548	13.032	8.1598
Ste†·	0.50	5.646	0.1857	0.0127	0.347	0.0004	0.0514	0.0781	0.3093	0.0257	0.106	0.1092	0.0886	0.0002	1.3564	0.1315	0.1679	1.0544	0.2750
	0.75	10.26	0.3050	0.0266	0.5891	0.0008	0.091	0.1415	0.7021	0.0962	0.2045	0.1716	0.1517	0.0003	2.5835	0.2286	0.2888	1.3932	0.4361
	0.90	16.70	0.4444	0.0486	1.1136	0.0013	0.1568	0.2366	1.1837	0.2789	0.4269	0.4579	0.2649	0.0005	4.087	0.3952	0.4550	1.8053	1.0289
	0.99	92.14	2.5360	0.1724	3.7413	0.0051	0.8643	0.2950	4.0343	1.2875	2.9488	3.5667	1.047	0.0022	16.856	1.2367	1.9927	4.2930	3.3823
Status	0.50	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
	0.75	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
	0.90	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
	0.99	NS	NS	S	S	NS	NS	NS	S	NS	NS	S	S	NS	NS	NS	S	NS	NS

† Note: Standard error (Ste) was estimated from the global quantile regression models.

4.3 Relationships between forest fires and environmental factors based on GWQR

The coefficient estimates of variables were computed based on GWQR and generally showed a similar change trend as the global QR models in terms of the mean or median of the model coefficients at different quantile levels (Table 3.5). GWQR produced the model coefficients for each location (pixel) across the study area at a specific quantile. Following Chen *et al*. (2012), we constructed the spatial maps of the GWQR model coefficients where the local t-test exceed ±1.96 (i.e., statistically significant) for each predictor variable at different quantile levels. Figures 3.5-3.8 illustrates the spatial map for the significant model coefficients of predictor variables across the study area at the four quantiles.

Figure 3.5(a)-(d) indicated that the local coefficient of elevation was statistically significant across the most study area at τ = 0.5, 0.75, and 0.90 quantiles. For the τ = 0.99 quantile, however, the larger negative coefficients were clustered in the northwest region. Similarly, the larger negative coefficients of slope were significantly concentrated on the north and west at τ = 0.50, 0.75, and 0.90 quantiles. Different from elevation and slope, the significant coefficients of aspect index were less and only located in the south of the study area (Figure 3.5(i)-(l)).

Greater positive coefficients of precipitation were located at the south at τ = 0.50, 0.75, and 0.90 quantiles. For τ = 0.99 quantile, the coefficients of precipitation became all negative and clustered in the north region. The positive significant coefficients of humidity were clustered at the northeastern and west regions, but humidity became less important for τ = 0.99 quantile. Unlike humidity, the more significant coefficients of temperature were concentrated in the northeastern coastal region when the quantile increased from 0.50 to 0.99 (Figure 3.6).

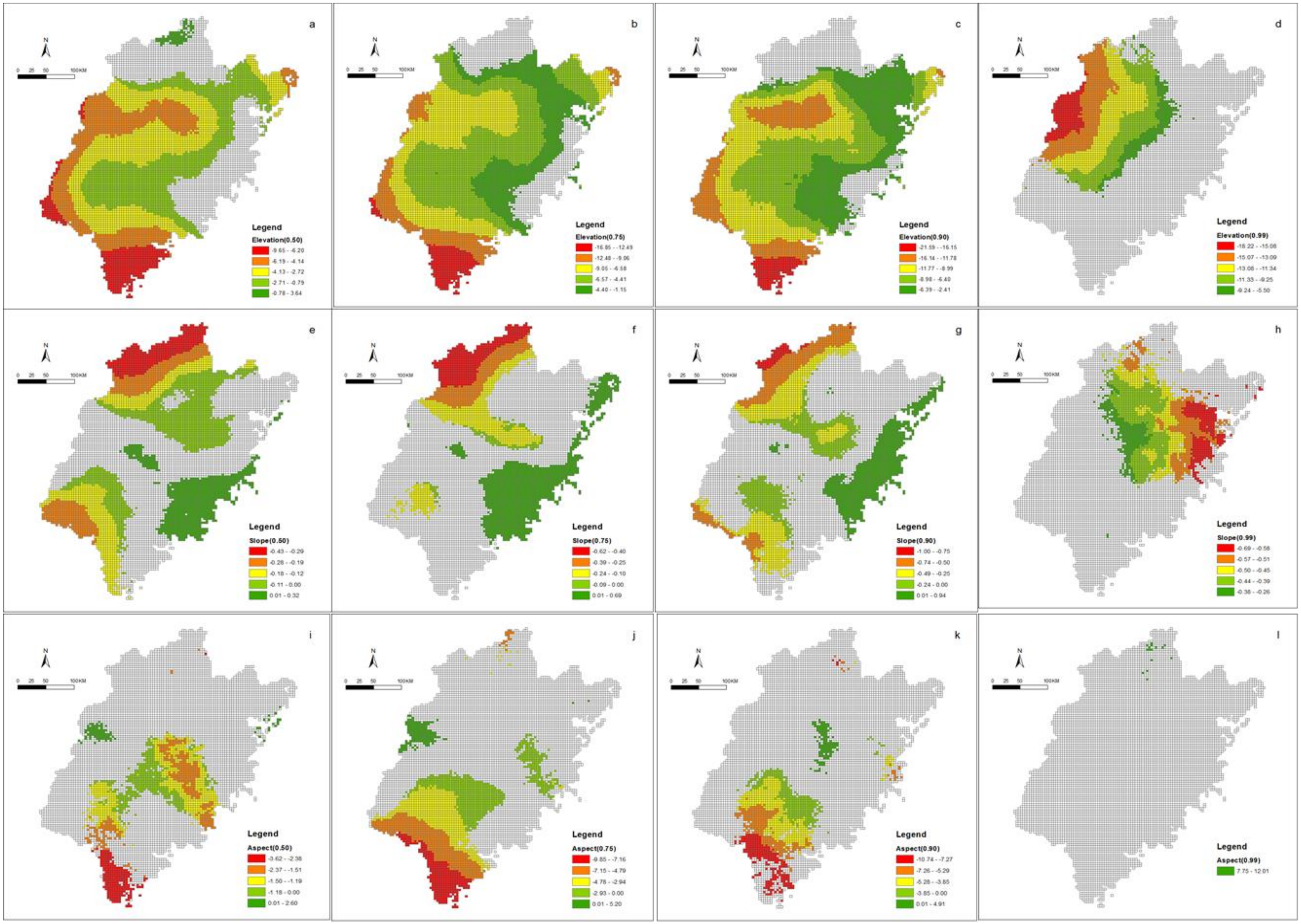


Figure 3.5 Spatial maps of the significant model coefficients of GWQR for topographical predictors.

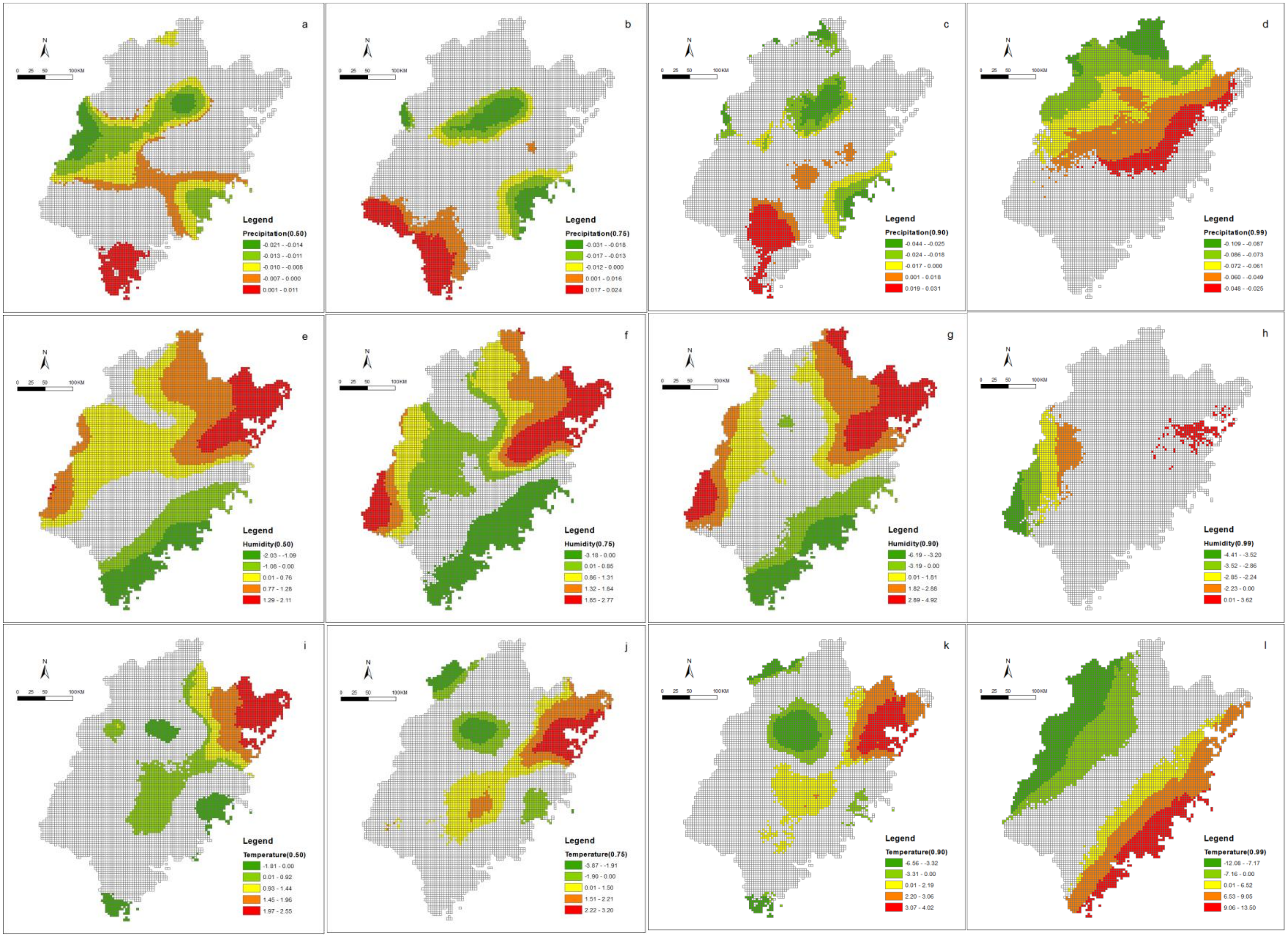
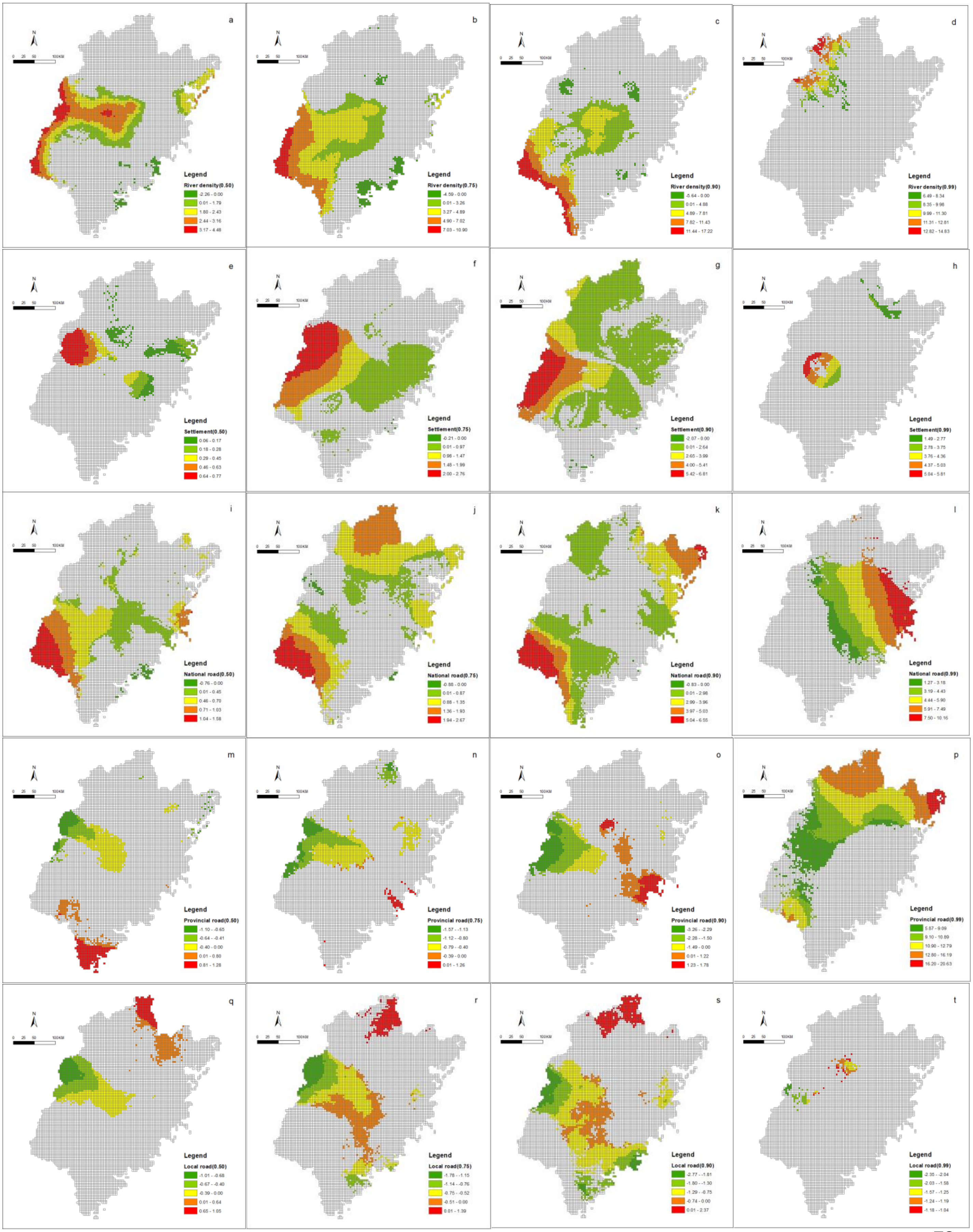


Figure 3.6 Spatial maps of the significant coefficients of GWQR for meteorological predictors.


a
Legend
River density(0.50)
-2.26 - 0.00
0.01 - 1.79
1.80 - 2.43
2.44 - 3.16
3.17 - 4.48
b
Legend
River density(0.75)
-4.59 - 0.00
0.01 - 3.26
3.27 - 4.89
4.90 - 7.02
7.03 - 10.90
c
Legend
River density(0.90)
-5.64 - 0.00
0.01 - 4.88
4.89 - 7.81
7.82 - 11.43
11.44 - 17.22
d
Legend
River density(0.99)
6.49 - 8.34
8.35 - 9.98
9.99 - 11.30
11.31 - 12.81
12.82 - 14.83
e
Legend
Settlement(0.50)
0.06 - 0.17
0.18 - 0.28
0.29 - 0.45
0.46 - 0.63
0.64 - 0.77
f
Legend
Settlement(0.75)
-0.21 - 0.00
0.01 - 0.97
0.98 - 1.47
1.48 - 1.99
2.00 - 2.76
g
Legend
Settlement(0.90)
-2.07 - 0.00
0.01 - 2.64
2.65 - 3.99
4.00 - 5.41
5.42 - 6.81
h
Legend
Settlement(0.99)
1.49 - 2.77
2.78 - 3.75
3.76 - 4.36
4.37 - 5.03
5.04 - 5.81
i
Legend
National road(0.50)
-0.76 - 0.00
0.01 - 0.45
0.46 - 0.70
0.71 - 1.03
1.04 - 1.58
j
Legend
National road(0.75)
-0.88 - 0.00
0.01 - 0.87
0.88 - 1.35
1.36 - 1.93
1.94 - 2.67
k
Legend
National road(0.90)
-0.83 - 0.00
0.01 - 2.98
2.99 - 3.96
3.97 - 5.03
5.04 - 6.55
l
Legend
National road(0.99)
1.27 - 3.18
3.19 - 4.43
4.44 - 5.90
5.91 - 7.49
7.50 - 10.16
m
Legend
Provincial road(0.50)
-1.10 - -0.65
-0.64 - -0.41
-0.40 - 0.00
0.01 - 0.80
0.81 - 1.28
n
Legend
Provincial road(0.75)
-1.57 - -1.13
-1.12 - -0.80
-0.79 - -0.40
-0.39 - 0.00
0.01 - 1.26
o
Legend
Provincial road(0.90)
-3.26 - -2.29
-2.28 - -1.50
-1.49 - 0.00
0.01 - 1.22
1.23 - 1.78
p
Legend
Provincial road(0.99)
5.87 - 9.09
9.10 - 10.89
10.90 - 12.79
12.80 - 16.19
16.20 - 20.63
q
Legend
Local road(0.50)
-1.01 - -0.68
-0.67 - -0.40
-0.39 - 0.00
0.01 - 0.64
0.65 - 1.05
r
Legend
Local road(0.75)
-1.78 - -1.15
-1.14 - -0.76
-0.75 - -0.52
-0.51 - 0.00
0.01 - 1.39
s
Legend
Local road(0.90)
-2.77 - -1.81
-1.80 - -1.30
-1.29 - -0.75
-0.74 - 0.00
0.01 - 2.37
t
Legend
Local road(0.99)
-2.35 - -2.04
-2.03 - -1.58
-1.57 - -1.25
-1.24 - -1.19
-1.18 - -1.04

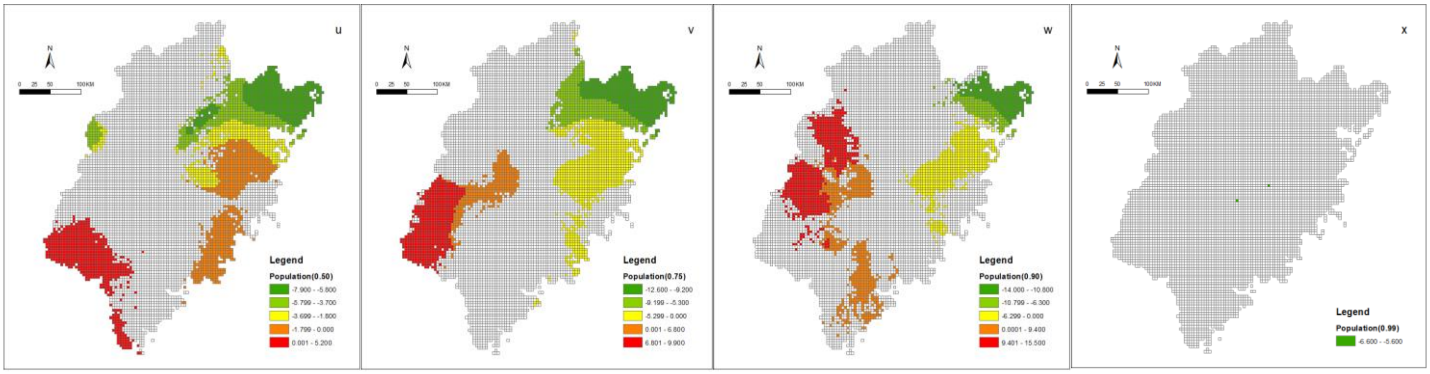


Figure 3.7 Spatial maps of the significant coefficients of GWQR for human related predictors.

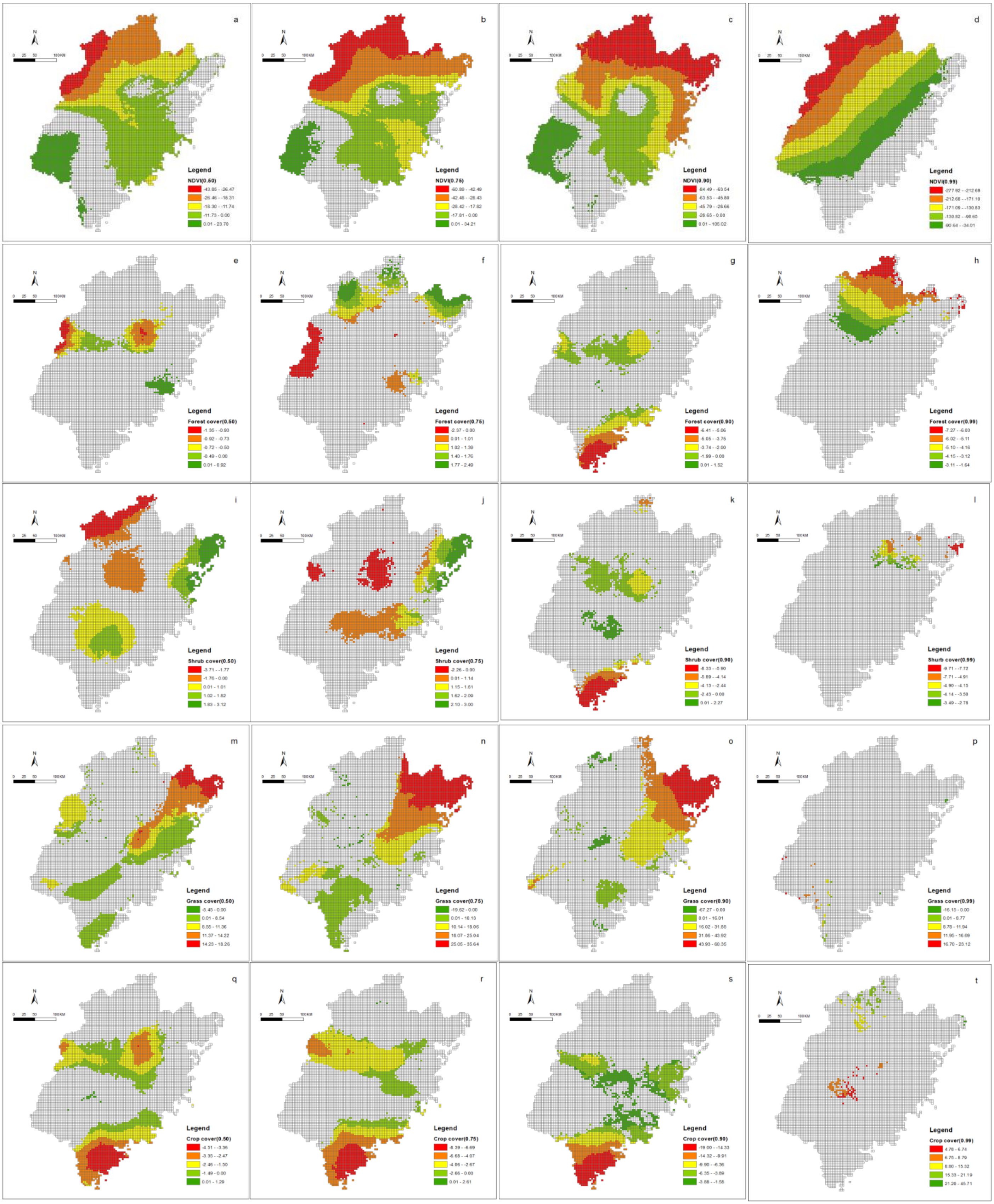


Figure 3.8 Spatial maps of the significant coefficients of GWQR for vegetation and land use predictors.

The significant coefficient of settlement density and national road density were clustered in the west regions. Similarly, the more significant coefficients of provincial road density were identified from south to north when the quantile increased from 0.50 to 0.99. The negative coefficients of population density were gathered along the eastern coast, while they became more positive in the western regions of the province for the quantile smaller than 0.90 (Figure 3.7).

Similarly, the significant negative coefficient of NDVI occupied most north and east regions of the province, but gradually became stronger from south to north. Forest cover and shrub cover in the global models were not statistically significant, but they showed significant impacts on the forest fire occurrence on in some particular locations of the study area at the four quantiles. For example, for τ = 0.50, 0.75, and 0.90 quantiles, grass cover was significant in the northeast and south, while crop cover was significant in the south and central regions (Figure 3.8).

4.4 Comparison on prediction accuracy between QR and GWQR

The comparison of the pinball loss values between global QR models and GWQR models were shown in Table 3.6 at the four quantiles. It revealed that the loss function values of GWQR models were all smaller than the corresponding global QR models at each quantile, clearly indicating that GWQR performed better than the global QR models. In addition, the improvement of GWQR forecasting increased as the quantile became larger. The pinball loss function values of GWQR reduced 3.74%, 6.37%, 9.20%, and 20.84% against the global QR models as the quantile increased from 0.50 to 0.99, respectively (Table 3.6).

Table 3.6 Pinball loss value for comparing GWQR against the global quantile models at different quantiles (smaller is better).

Model	τ = 0.50	τ = 0.75	τ = 0.90	τ = 0.99
Global Quantile	11716.58	12093.17	8622.89	2180.45
GWQR	11277.26	11322.69	7928.88	1726.00
Improvement	3.74%	6.37%	9.20%	20.84%

5. Discussion

Our results indicated that the importance of the predictor variables varied at different quantiles of forest fire occurrence. For example, the aspect index was statistically significant at a relatively low quantile ($\tau = 0.50$), but not significant at the higher quantiles ($\tau = 0.90$ and 0.99). In contrast, the grass cover was less associated with forest fires at a lower quantile of fire occurrence ($\tau = 0.50$), while significantly related to the upper quantiles (e.g., $\tau = 0.75$ and 0.90) of forest fires (Table 3.3). In practice, when the upper or lower quantiles of the response variable are more interested to researchers or policy makers, the QR models are particular useful for exploring the full range of the variable relationships (Cade and Noon 2003). However, QR only offers a picture of global regularity. The issues of local spatial constituents, structure, variability, and complexity are still undetected.

Methodologically, GWQR is extended from global QR, which can account for both spatial effects and quantile distributions of the response variable and make no distributional assumption about the error term of the model. In other words, GWQR is not stuck to the issues of violating the assumption of an ordinary least squares (OLS) model such as normality and constant variance. For example, in this study, about 20% of the grid pixels did not have forest fires. Thus, the overdispersion of the response variable violates both assumptions of normality and constant variance, including both global and GWR. Therefore, GWQR is unbiased when the tails and central location of the conditional distribution vary differently in the response variable (Chen *et al*. 2012). Specifically, GWQR allows exploring non-stationarity across the study area depending on the various quantiles of the response variable, which are particularly important in the fields of forest fire risk assessment. For example, it indicated that slope had important negative influence on the fire occurrence at $\tau = 0.50$, 0.75, and 0.90 quantiles on the north and west edge of the study area where the terrains were relatively flatter than other areas (Figure 3.5). However, the difference of 9% HON pixels ($\tau = 0.90$ and 0.99) were dramatically changed the relationships between slope and fire occurrence (the significant coefficient was concentrated in the central regions of the study area). Similarly, the significant positive coefficients of precipitation and humidity were detected in some certain locations of the study area where daily rainfall around 1600 mm (or less) and relative humidity is less than 77% (Figure 3.6) at $\tau = 0.50$, 0.75, and 0.90 quantiles. While τ increased to $\tau = 0.99$ quantile, both precipitation and humidity became less important to the fire occurrence. The findings suggested that a few HON events may dramatically affect the entire distribution of the relationships between fire occurrence and a specific environmental factor. Ignoring this important fact would mislead the fire regimes analysis and management.

In addition, our results also indicated that the estimated model coefficients of all predictor variables were very different between the model of $\tau = 0.50$ quantile and the model of $\tau = 0.99$ quantile (Table 3.3). It was noticeable that not only the model coefficients were dissimilar at different quantiles of forest fires, the locations of significant coefficient of each predictor were also divergent (Figures 3.5-3.8), indicating that the impacts of these environmental factors were spatially varying at different quantiles of forest fires. Therefore, GWQR is able to help researchers to explore the locally detailed relationships simultaneously across different conditional distributions of the response variable. Ultimately, those deeper insights may help planning and investment into management activities.

Overall, our findings on the relationships between fire occurrence and specific environmental factors were consistent with the previous studies that applied different method in the Fujian province in terms of the coefficient signs (+ or -) of factors (Guo *et al*. 2016a, 2017). In this study, both QR and GWQR indicated that precipitation and relative humidity had positive impacts on fire occurrence across the study area at different quantiles ($\tau = 0.50$, 0.75, and 0.90). Although these findings seem contradictory to the general understanding that as rainfall increases, fewer fires will occur, one explanation is that more rainfall and humidity are beneficial to the growth of ground-cover vegetation, and increased amounts of surface fuel load, which increases the risk of forest fire occurrence. In the

Kruger National Park of South Africa, van Wilgen *et al.* (2000) observed a strong positive correlation between precipitation rates and fire activity. Spessa *et al*. (2005) and Randerson *et al*. (2005) also found a similar positive association between precipitation and fire activity in north Australia using different satellite data sets.

As an improved regression method toward a spatial quantile-based analysis, some issues of GWQR still need to be further developed and discussed. One issue is the bandwidth selection. In this study, the variogram model (Bailey and Gaterell 1995) was fitted through the observed response variable and residuals based on the global quantile models. Given the difference of optimal model and estimated bandwidths (Table 3.4), we chose the bandwidth referenced on the residuals of global quantile models at the four quantiles for the further GWQR analysis. However, the optimal bandwidth may be selected in terms of other criterions, such as subjective and smallest cross-validation error (Chen *et al*. 2012).

6. Conclusion

In this study we applied global QR and GWQR to model the relationships between forest fire occurrence and environmental factors at different quantiles (τ = 0.50, 0.75, 0.90, and 0.99). A total of seventeen (17) predictor variables from four categories (topography, meteorology, human, vegetation coverage, and land use) were collected for fitting both global and GWR quantile regression models.

Our results showed that the impacts of those environmental factors on forest fires significantly varied not only at different quantiles of fire occurrence, but also across the geographical study area. Some driving factors, such as elevation, slope, NDVI, settlement density, national road density, and grass cover, were statistically significant at the four quantiles. The degree of the significance, however, varied across different regions of the study area between different quantiles and /or within the same quantile. Other factors were statistically significant at particular quantiles (e.g., either lower quantiles or higher quantiles). Even few factors were not important in the global QR models, their model coefficients might be important in particular regions.

It was evidence that GWQR integrated the entire distributions (i.e., different quantiles) of forest fires and the spatially variations of the relationships between forest fires and driving factors. Thus, GWQR would provide useful information on the different levels of forest fire risks, as well as high risky locations of forest fires across the study area. Hopefully, the information would assist the government agencies to make better decisions on where and what the fire management and fire prevention should be focused on in order to reduce economic expenses and improve the efficiency of forest fire management.

7. References

Bailey T.C., Gatrell A.C., 1995. Interactive spatial data analysis. Longman Scientific and Technical, Routledge.

Barros A.M., Pereira J.M.C., 2014. Wildfire selectivity for land cover type: Does size matter? Public Library of Science ONE 9(1): e84760. Available at: https://doi.org/10.1371/journal.pone.0084760

Blanchi R., Jappiot M., Alexandrian D., 2002. Forest fire risk assessment and cartography - A methodological approach. *Proceedings of IV International Conference on Forest Fire Research,* November 18-23, 2002. Luso, Portugal.

Bradshaw L.S., Deeming J.E., Burgan R.E., Cohen J.D., 1984. The 1978 National Fire-Danger Rating System: Technical documentation. U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station General Tech. Rep. INT-169, 44 pp.

Brillinger D.R., Preisler H.K., Benoit J.W., 2003. Risk assessment: A forest fire example. In Science and statistics: A Festschrift for Terry Speed. (Ed. DR Goldstein). pp. 177-196. (Institute of Mathematical Statistics: Beachwood, OH)

Burgan R., Klaver R.W., Klaver J.M., 1998. Fuel models and potential from satellite and surface observations. *International Journal of Remote Sensing* 8: 159-170.

Cade B.S., Noon B.R., 2003. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* 1(8): 412-420.

Calkin D.E., Cohen J.D., Finney M.A., Thompson M.P., 2014. How risk management can prevent future wildfire disasters in the wildland-urban interface. *Proceedings of the National Academy of Sciences of the United States of America* 111(2): 746-751.

Chen C., Wei Y., 2005. Computational issues for quantile regression. *Sankhyā: Indian Journal of Statistics* 67: 399-417.

Chen V.Y.J., Deng, W.S., Yang, T., Matthews, S.A., 2012. Geographically weighted quantile regression (GWQR): An application to U.S. mortality data. *Geographical Analysis* 44: 134-150.

Finney M.A., McHugh C.W., Grenfell J.C., Riley K.L., Short K.C., 2011. A simulation of probabilistic wildfire risk components for the continental United States. *Stochastic Environmental Research and Risk Assessment* 25: 973-1000.

Foody G.M., 2003. Geographical weighting as a further refinement to regression modeling: an example focused on the NDVI-rainfall relationship. *Remote Sensing of Environment* 88(3): 283-293.

Fotheringham A.S., Brunsdon C., Charlton M., 2002. Geographically weighted regression. Wiley, New York. 284 p.

Fotheringham A.S., Charlton M.E., Brunsdon C., 1998. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environmental and Planning A* 30(11): 1905-1927.

Guo F., Innes L.J., Wang G., Ma X., Sun L., Hu H., Su Z., 2015. Historic distribution and driving factors of human-caused fires in the Chinese boreal forest between 1972 and 2005. *Journal of Plant Ecology* 8(5): 480-490.

Guo F., Ju Y., Wang G., Alvarado E.C., Yang X., Ma Y., Liu A., 2018. Inorganic chemical composition of PM2.5 emissions from the combustion of six main tree species in subtropical China. *Atmospheric Environment* 189: 107-115.

Guo F., Wang G., Su Z., Liang H., Wang W., Lin F., Liu A., 2016a. What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire* 25: 505-519.

Guo F., Selvalakshmi S., Lin F., Wang G., Wang W., Su Z., 2016b. Geospatial information on geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese boreal forest. *Canadian Journal of Forest Research* 46: 582-594.

Guo F., Zhang L., Jin S., Tigabu M., Su, Z., Wang W., 2016c. Modeling anthropogenic fire occurrence in the boreal forest of China using logistic regression and Random Forest. *Forests* 7: 250-264.

Guo F., Su Z., Wang G., Sun L., Tigabu M., Yang X., Hu H., 2017. Understanding fire drivers and relative impacts in different Chinese forest ecosystems. *Science of the Total Environment* 605-606: 411-425.

Guo L., Ma Z., Zhang L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. *Canadian Journal of Forest Research* 38(9): 2526-2534.

Hoyo L.V.D., Pilar Martín Isabel M., Javier Martínez Vega F., 2011. Logistic regression models for human-caused wildfire risk estimation: analysing the effect of the spatial accuracy in fire occurrence data. *European Journal of Forest Research* 130(6): 983-996.

Huang Q., Zhang H., Chen J., He M., 2017. Quantile regression models and their applications: A review. *Journal of Biometrics and Biostatistics* 8(3): 354-359.

Hutchinson, 2004. ANUSPLIN Version 4.3. Centre for Resource and Environmental Studies, Australian National University, November 23, 2005. Available at http://cres.anu.edu.au/outputs/anusplin.php

Koenker R., 2005. Quantile Regression. Cambridge University Press. Cambridge, UK.

Koenker R., Bassett J.R.G., 1978. Regression quantiles. *Econometrical. Journal of the Econometric Society* 46: 33-50.

Koutsias N., Martínez-Fernández J., Allgöwer B., 2010. Do factors causing wildfires vary in space? Evidence from Geographically Weighted Regression. *GIScience and Remote Sensing* 47: 221-240.

Liu Z., Yang J., Chang Y., Weisberg P. J., He H. S., 2012. Spatial patterns and drivers of fire occurrence and its future trend under climate change in a boreal forest of northeast China. *Global Change Biology* 18: 2041-2056.

Lopez A.S., San-Miguel-Ayanz, J., Burgan R., 2002. Integration of satellite sensor data, fuel type maps and meteorological observations for evaluation of forest fire risk at the pan-European scale. *International Journal of Remote Sensing* 23: 2713-2719.

Mandallaz D., Ye R., 1997. Prediction of forest fires with Poisson models. *Canadian Journal of Forest Research* 27: 1685-1694.

Martínez-Fernández J., Chuvieco E., Koutsias N., 2013. Modeling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression. *Natural Hazards and Earth System Sciences* 13: 311-327.

McCoy V.M., Burn C.R., 2005. Potential alteration by climate change for forest-fire regime in the boreal forest of central Yukon Territory. *Arctic* 58: 276-285.

Mckenney D.W., Pedlar J.H., Papadopol P., Hutchinson M.F., 2006. The development of 1901–2000 historical monthly climate models for Canada and the United States. *Agricultural and Forest Meteorology* 138(1-4): 69-81.

McKenzie D., Peterson D.L., Agee J.K., 2000. Fire frequency in the Interior Columbia Building regional models from fire history data. *Ecological Applications* 10(5): 1497-1516.

Moran P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika* 37(1): 17-23.

Mueller J.M., Loomis J.B., 2014. Does the estimated impact of wildfires vary with the housing price distribution? A quantile regression approach. *Land Use Policy* 41: 121-127.

Nakaya T., Fotheringham A.S., Brunsdon C., Charlton M., 2005. Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine* 24: 2695-2717.

Oliveira S., Oehler F., San-Miguel-Ayanz J., Camia A., Pereira J., 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and Random Forest. *Forest Ecology and Management* 275: 117-129.

Oliveira, S., Pereira, J.M.C., San-Miguel-Ayanz, J., Lourenço, L., 2014. Exploring the spatial patterns of fire density in Southern Europe using geographically weighted regression. *Applied Geography* 51: 143-157.

Preisler H.K., Brillinger D.R., Burgan R.E., Benoit J.W., 2004. Probability based models for estimation of wildfire risk. *International Journal of Wildland Fire* 13(2): 133-142.

Preisler H.K., Westerling A.L., 2007. Statistical model for forecasting monthly large wildfire events in western United States. *Journal of Applied Meteorology and Climatology* 46: 1020-1030.

Rijal B., 2018. Quantile regression: An alternative approach to modeling forest area burned by individual fires. *International Journal of Wildland Fire* 27(8): 538-549.

Rodrigues M., de la Riva J., Fotheringham S., 2014. Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. *Applied Geography* 48: 52-63.

Rodrigues, M., Jiménez-Ruano, A., Peña-Angulo, D., de la Riva, J., 2018. A comprehensive spatial-temporal analysis of driving factors of human-caused wildfires in Spain using geographically weighted Logistic regression. *Journal of Environment Management* 225: 177-192.

SAS Institute, Inc., 2013. STAT 9.4 Users' Manual. SAS Institute, Inc. Cary, USA.

Scholze M., Knorr W., Arnell N.W., Prentice I.C., 2006. A climate-change risk analysis for world ecosystems. *Proceedings of the National Academy of Science of the United States of America* 103: 13116-13120.

Scott J.H., Thompson M.P., Calkin D.E., 2013. A wildfire risk assessment framework for land and resource management. Gen. Tech. Rep. RMRS-GTR-315. US Forest Service Rocky Mountain Research Station, Fort Collins, CO.

Spessa A., McBeth B., Prentice C., 2005. Relationships among fire frequency, rainfall and vegetation patterns in the wet-dry tropics of northern Australia: An analysis based on NOAA-AVHRR data. *Global Ecology and Biogeography* 14(5): 439-454.

Su Z., Hu H., Tigabu M., Wang G., Zeng A., Guo F., 2019. Geographically weighted negative binomial regression model predicts wildfire occurrence in the Great Xing'an Mountains better than negative binomial model. *Forests* 10: 377.

Syphard A.D., Radeloff V.C., Keeley J.E., Hawbaker T.J., Clayton M.K., Stewart S.I., Hammer R.B., 2007. Human influence on California fire regimes. *Ecological Applications* 17(5): 1388-1402.

van Wilgen B.W., Biggs H.C., O'Regan S.P., Mare N., 2000. A fire history of the savanna ecosystems in the Kruger National Park, South Africa, between 1941 and 1996. *South Africa Journal of Science* 96(4): 167-178.

Wu Z., He H.S., Yang J., Liu Z., Liang, Y., 2014. Relative effects of climatic and local factors on fire occurrence in boreal forest landscapes of northeastern China. *Science of the Total Environment* 493: 472-480.

Wybo J.L., Guarnieri F., Richard B., 1995. Forest fire danger assessment methods and decision support. *Safety Science* 20: 61-70.

Xystrakis F., Koutsias N., 2013. Differences of fire activity and their underlying factors among vegetation formations in Greece. *iForest-Biogeosciences and Forestry* 6: 132-140.

Yu K., Lu Z., Stander J., 2003. Quantile regression: Applications and current research areas. *Journal of Royal Statistical Society* 52(3): 331-350.

Yu, L., Yang, Z., Tang L., 2018. Quantile estimators with orthogonal pinball loss function. *Journal of Forecasting* 37(3): 401-417.

Zhang L., Bi H., Gove J.H., Heath L.S., 2005. A comparison of alternative methods for estimating the self-thinning boundary line. *Canadian Journal of Forest Research* 35: 1507-1514.

Zhang X., Kang S., Zhang L., Liu J., 2010. Spatial variation of climatology monthly crop reference evapotranspiration and sensitivity coefficients in Shiyang river basin of northwest China. *Agricultural Water Management* 97(10): 1506-1516.

Chapter IV: Studying the Relationship between Proportion of Forest Fire Occurrence and Environmental Factors Using Beta Regression

Abstract. When a response variable is binary (i.e., yes or no), logistic regression is usually applied to model the probability of the event. However, when a response variable is proportion / percentage / rate, beta regression is more appropriate for providing direct and quantitative views on the relationships between the response variable and explanatory variables. In this study, we applied the global beta regression, geographically weighted beta regression (GWBR), and classical geographically weighted regression (GWR) to explore the spatially varying relationships between the proportion of forest fire occurrence and topographical, meteorological, human, vegetation coverage, and land cover factors in Fujian province, southeast China. Our results indicated that, in general, the proportion of forest fires was higher in lower elevation, stronger sunshine, denser settlement, and less cropland coverage. Environmental factors were spatially variedly related to the proportion of fire occurrence in the study region. With regards to model fitting and predicting, the global beta and GWBR models exhibited better performances than the classical GWR model for the proportion or rate response variable. Additionally, the GWBR model was productive at targeting the essential hotspots of predictor variables. Therefore, GWBR is an appropriate method to analyze the probability of forest fire occurrence which spatially varied and within range between (0, 1), and support better understanding for local prevention and management of forest fire.

Keywords: proportion of forest fire occurrence, beta regression, geographically weighted beta regression

1. Introduction

Forest fires represent a significant threat to ecological and social systems (Pyne *et al*. 1996) causing escalating social, eco-environmental and fiscal costs along with the changing climate (Calkin *et al*. 2014; Westerling *et al.* 2006). Every year in China, about 10,000 wildfires occurred with 820,000 hectares burnt areas (Guo *et al.* 2015). Despite its threats to people's lives, infrastructures, and valuable environmental resources, wildfire plays a natural function in forest renewal and succession (Podur *et al.* 2003; Chang *et al.* 2007). To reduce the losses caused by forest fires, prevention and suppression are the main tasks of forest management agencies, including better understanding of the occurrence probability and patterns of forest fires, the impacts of environmental drivers on forest fire ignition, develop, and spread, and efficient and effective strategies for forest fire prevention, detection, control, and suppression.

To investigate the forest fire ignition probability, various methods and models have been developed. In the review of mathematical models of wildfires since 1940 (Pastor *et al.* 2003), the empirical model, which applied the statistical methods to discover how the impact factors influenced forest fires, was considered one of three major methods in general. Within the statistical methods, logistic regression is the most popular technique because it is reasonably flexible and accepts a mixture of continuous and categorical variables, as well as non-normally distributed variables (Catry *et al.* 2009). It is used to quantify the relationships and predictions between the occurrence probability of forest fires and potential explanatory variables. Martell *et al.* (1987) developed a procedure based on a logistic model for predicting daily people-caused forest fire occurrences in Ontario. Vega Garcia *et al.* (1995) used it to predict the number of fire-days in the Whitecourt Forest of Albert, Canada. In 2004, Preisler *et al.* presented a probability-based model for estimating fire risk, by fitting a spatially and temporally explicit non-parametric logistic regression to grouped data. Then, Preisler and Westerling (2007) described a method based on the logistic regression to accommodate the relationships between fire-danger predictors and the probability of large fire events. Lozano *et al*. (2007) assessed the performance of several spectral indices derived from Landsat data when modeling fire occurrence probability through logistic regression. Hoyo *et al.* (2011) estimated human-caused wildfire risk based on the logistic regression in Spain. Chang *et al.* (2013) predicted the fire occurrence pattern with the logistic regression in Heilongjiang Province, China. *Guo et al.* (2016) used the logistic regression to identify the drivers of wildfire and predict the likelihood of fire occurrence in southeast China.

Different from the feature of logistic regression that transfers a binary response variable to an odds ratio, beta regression is an alternative technique to model probability, proportion, or rate. It is introduced by Ferrari and Cribari-Neto (2004), based on the assumption that the response variable follows a beta distribution, which is a family of continuous probability distributions strictly defined on the interval (0, 1) with two shape parameters. Those two positive shape parameters control the shape of distribution within the (0, 1) interval. Therefore, the beta distribution is very flexible for continuous response variables in an (0, 1) interval with two shape parameters. However, the previous application of beta distribution did not involve the situation that the response variable can be modeled as a function of exogenous variables until beta regression was proposed (Ferrari and Cribari-Neto 2004).

By transferring two original shape parameters into mean and dispersion, the beta regression model is interpretable in terms of the mean of the response variable (proportion or rate). The function of mean is given by a linear predictor defined by regression coefficients and explanatory variables. Thus, the beta regression model is related to other variables through a regression structure, similar to generalized linear models. The parameter estimation is performed by maximum likelihood (more details of beta regression will be represented in the method section). Therefore, the beta regression model is more suitable for modeling continuous response variables such as probability, percentages, proportions, rates, and fractions without data transformation.

Because of its flexibility and empirical application, more beta regression models have been developed, extending to fit different conditions (Ferrari 2013). For instance, time-series beta regression model for periodic data with a trend by Rydlewski in 2007; inflated beta regression model (Ospina and Ferrari 2012) considering extreme values of zero and one; semi-parametric beta regression (Branscum *et al*. 2007) based on Bayesian inference method to estimate household expenditure and genetic distance between foot-and-mouth disease viruses; multivariate beta regression (Souza and Moura 2012) for jointly modeling two or more than two variables whose values belong to the (0, 1) interval; mixed beta regression (Zimprich 2010) including the random effect for longitudinal data; beta rectangular regression model (Bayes *et al.* 2012) flexible for outlying observations. The literature discussed on special topics of diagnostics, tests, robust inference, optimal designs, and maximum likelihood estimators in beta regression has been growing over the last few years. To date, beta regression models are applied to a wide range of study fields, such as medicine, deontology, hydrobiology, economics, aquaculture nutrition, forest sciences, education, political science, waste management, and etc. (Ferrari 2013). In the field of the forest fire science, Ríos-Pena *et al.* (2018) proposed the Zero-One-Inflated structured additive beta regression to study wildfire occurrence and burnt area simultaneously in Spain.

To our best knowledge, there is no study to model the proportion of forest fires using beta regression. Therefore, we propose to apply the modeling technique to estimate the relationships between forest fire proportion or probability and potential environmental and human factors. However, it is well known that spatial autocorrelation and heterogeneity exist across forest ecosystems, and the relationships between forest fires and environmental factors are spatially nonstationary. In addition to the global beta regression model, we developed the geographically weighted beta regression (GWBR) to explore and quantify the spatially varying association across the study area. The GWBR model was designed referring to the framework of Geographically Weighted Regression (GWR) defined by

Fotheringham *et al.* (1996), Brunsdon *et al.* (1996, 1998), and Fotheringham *et al.* (2002), which is a spatial modeling technique that takes non-stationary variables into consideration and investigate the local relationships between these predictors and an outcome of interest. GWBR is the extension d of beta regression concepts to GWR for locally modeling data within the interval of (0, 1).

The objectives of this research were using beta regression: (1) to identify potential driving factors of forest fire proportion, from the infrastructure, topography, meteorology, human activity, and land coverage; (2) to explore spatial variability relationships between the probability of forest fire occurrence and influential risk factors; (3) to target the localized significant explanatory variables for prediction and prevention of forest fires; and (4) to compare the performances of global beta, GWBR, and GWR for modeling the proportion of forest fire.

2. Data

2.1 Study area

Fujian province is located in a sub-tropical region of China and has a total land area of 124,000 km^2 (Figure 4.1). It ranks the highest forest coverage in the nation, with about 66% of Fujian province covered by forests and vegetation. It also experiences high annual forest fire incidences, with nearly 15,000 forest fires occurring from 2000 to 2010 (Guo *et al*. 2018). The dominant tree species in Fujian include Massoniana (*Pinus massoniana* Lamb.), Chinese fir (*Cunninghamia lanceolate* (Lamb.) Hook), Casuarina (*Casuarina equisetifolia* L.), and Pubescens (*Phyllostachys heterocycle* (Carr.) Mitford cv. *Pubescens*). The climate is warm and humid with an average annual rainfall of 1400 – 2000 mm and average temperature of 17 – 21 °C. Forest fire season typically spans from September to April (Guo *et al*. 2016a).

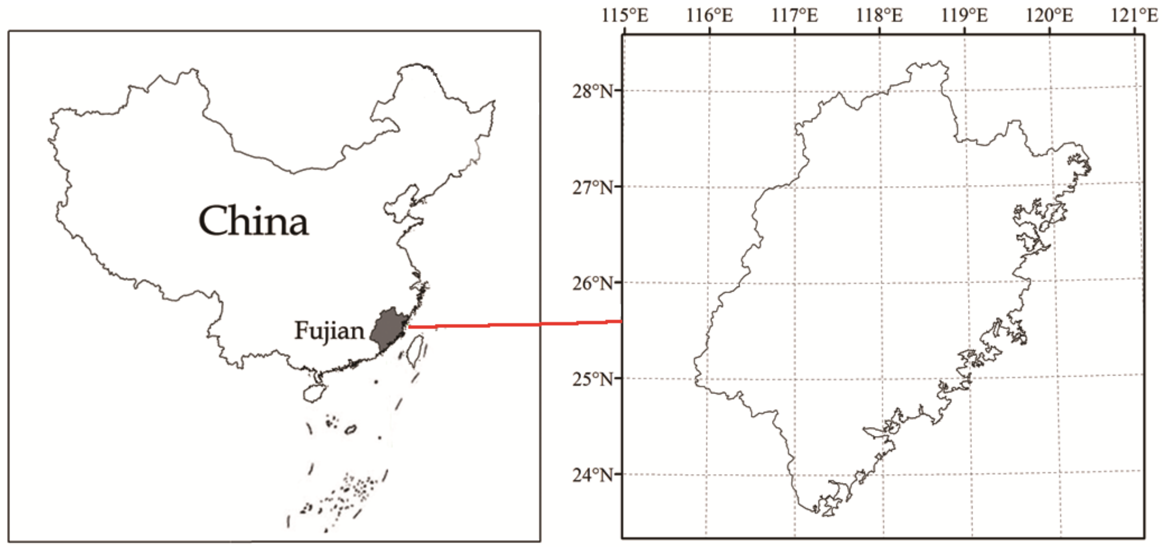


Figure 4.1 Location map of Fujian Province.

2.2 Data collection

2.2.1 Fire data (response variable)

In this study, we used MODIS hotspots product (MOD14A1) which has been considered as a reliable and suitable source for monitoring forest fires to analyze the relationships between forest fire occurrence and environmental factors in Fujian, China (Guo *et al*. 2016 a,b; Su *et al*. 2019). The time period of the study is 16 years (2001 - 2016). Since MOD14A1 cannot distinguish forest fires from non-forest fires that occur in cities/towns, construction sites, agricultural lands, and other areas, we further processed the fire data by: (1) removing the fire points in cities/towns, construction sites, and farmland based on an 1 km resolution land-use map; and (2) extracting fire points based on the time of fire occurrence within the fire season of the study area (September 15 to April 30 of the following year). The whole study area was divided into 4 × 4 km grids (a total of 7433 grids) using ArcGIS 10.2 (ESRI, 2010) and the total number of forest fire occurrences in each grid was calculated. Using the information of forest fire counts and years, we calculate the proportion of forest fire occurrence during the 16-year period (2001 - 2016) by,

$P = N / 16 + W$

where P is the proportion of forest fire occurrence, N represents the total number of years that each grid had wildfire occurred, sixteen is the total years, and W is the weight probability of annual extra wildfire counts as:

$$W = (T - N) * (10\% / 153)$$

where T is the total number of forest fire in that grid for the 16-year period. With the principle that the probability does not exceed 100%, the maximum weight probability of extra fire occurrence is 10%. For each extra wildfire occurrence, the weight is calculated as the 10% divided by 153, which is the maximum extra wildfire number in this study. Statistic summary of the proportion of forest fire occurrence is listed in Table 4.1, and its frequency distribution and geospatial map are presented in Figures 4.2 and 4.3(a), respectively.

Table 4.1 Descriptive statistics of response and predictor variables and correlation coefficients to response variable.

Variable	Mean	Median	Std Dev	Minimum	Maximum	Correlation coefficient
Proportion of Fire Occurrence	0.1567	0.1250	0.1497	0.0001	0.9999	1.00
Elevation (km)	0.4885	0.4688	0.2714	-0.0064	1.7577	-0.304
Slope (degree)	19.78	20.72	5.77	0.105	37.68	-0.214
Aspect Index	-0.0073	-0.0116	0.1186	-0.5728	0.5523	-0.011
Precipitation (mm/day)	1670	1682	158.1292	1247	2042	0.019
Temperature (°C)	18.30	18.00	1.16	15.27	21.12	0.036
Humidity (%)	76.07	76.00	1.4118	73.00	79.20	0.072
River Density (%)	0.5969	0.5584	0.1962	0.0015	1.6537	0.100
Settlement Density (%)	0.4523	0.0000	3.7411	0.0000	95.4832	0.091
National Road Density (%)	0.2006	0.0000	0.7451	0.0000	6.5424	0.130
Provincial Road Density (%)	0.2176	0.0000	0.5220	0.0000	5.3110	0.083
Local Road Density (%)	1.052	1.051	0.5848	0.0000	4.873	0.005
Population (people/km^2)	283.9	133.5	535.7	56.8	17899.5	-0.001
NDVI	0.7742	0.7910	0.0662	0.3287	0.8729	-0.139
Forest Cover (%)	51.752	55.374	39.487	0.000	100.00	-0.044
Shrub Cover (%)	18.608	0.000	28.160	0.000	99.99	-0.044
Grass Cover (%)	1.409	0.000	9.583	0.000	99.99	0.043
Crop Cover (%)	10.161	0.000	22.499	0.000	99.99	0.071

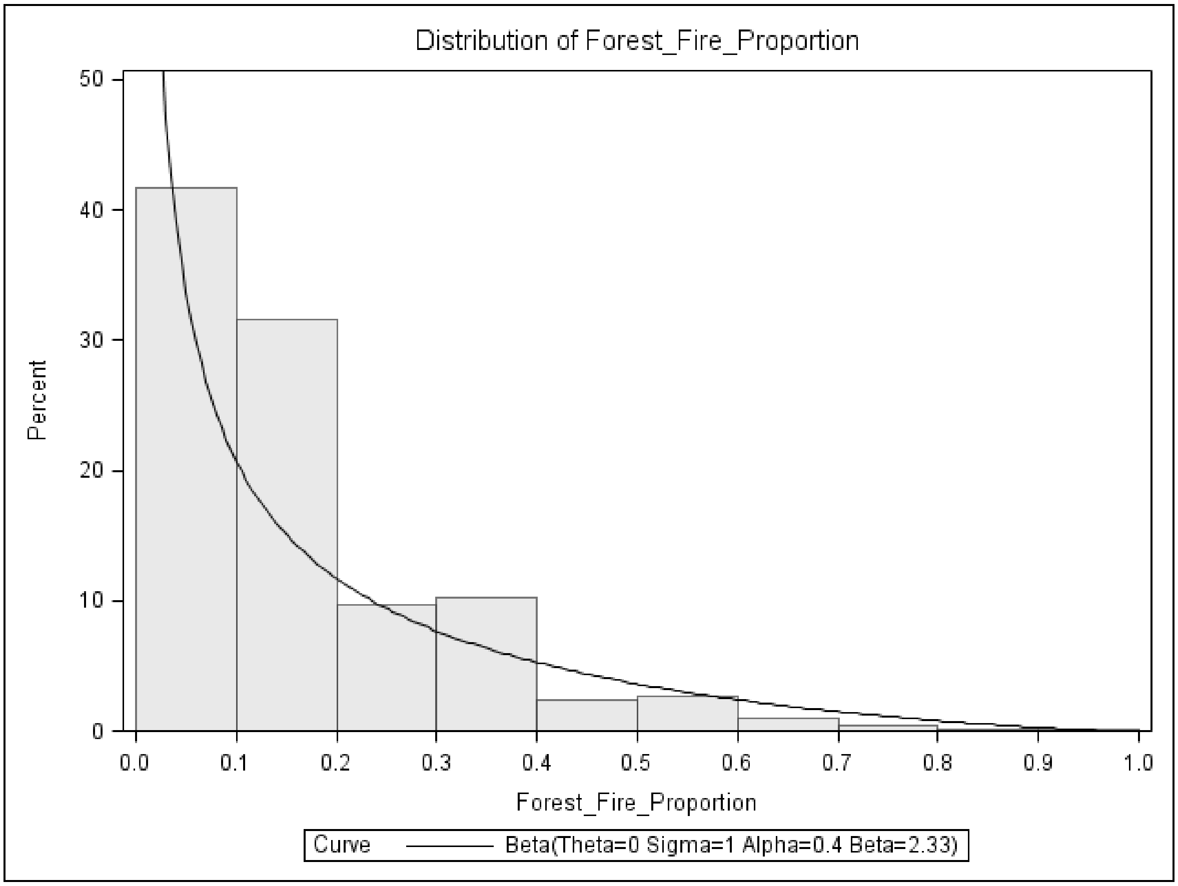


Figure 4.2 Frequency distribution of forest fire proportions with fitted beta curve.

2.2.2 Potential driving factors (predictor variables)

A total of 18 predictor or explanatory variables were collected and grouped into four categories, including topographical, meteorological, human related, and vegetation and land use predictors. The specific variable collection processes were as follows:

Topographic variables

Topographic variables included elevation (km), slope (degree), and aspect index. High resolution (25 m) Digital Elevation Model (DEM) data were collected from the National Administration of Surveying, Mapping and Geoinformation of China (http://www.gscloud.cn/sources). The slope and slope direction were derived from DEM, and aspect was then converted into an aspect index using the following formula: Aspect Index = $\cos(\theta \times \pi / 180)$, where θ is the degree of slope generated in ArcGIS ranging from 0 – 360° so that the aspect index ranges from -1 to 1 (Guo *et al.* 2017). The average elevation, slope, and aspect index of each grid were then extracted using ArcGIS 10.2.

Meteorological variables

Meteorological variables included precipitation (mm/day), temperature (°C), and relative humidity (%), which were obtained from the platform of National Earth System Science Data Center (http://www.geodata.cn), an important component of National Science and Technology Infrastructure. The climatic variables are interpolated from ANUSPLIN, a software package developed by Hutchinson (2004) based on the thin-plate smoothing method to generate hydrometeorological maps. ANUSPLIN includes a linear covariate to represent the elevation dependent meteorological factors, and it outperformed in climate interpolation (Zhang *et al*. 2010) and long period climatic data (McKenney *et al*. 2006). Raster calculator in ArcGIS 10.2 was used to calculate the annual average of each meteorological variable for each grid from year 2001 to 2016. Precipitation and temperature impact the occurrence of forest fire by limiting the fuel moisture content. Therefore, they are reasonable and effective alternative fuel factors when other fuel factor is not available. In addition, the annual meteorological factor is a traditional and better indicator to measure the influence of climate change on forest fires, compared to the average meteorological data during the period of forest fire (Scholze *et al*. 2006; McCoy and Burn 2005; Xystrakis *et al.* 2013).

Human factors

Human factors included the socioeconomic variables (per capita GDP and population density) and infrastructural variables. The GDP and population data were obtained from Resource and Environmental Data Cloud Platform (http://www.resdc.cn/Default.aspx) and the data resolution was 1 km. The infrastructural variables included road density (km/km^2, ratio of road length to the grid area) and water density. The 1:250,000 vector map of infrastructure was provided by the National Geomatics

Center of China (http://www.ngcc.cn/). We classified the road into national, provincial, and local road. Their buffer areas were built based on 50 m, 25 m, and 10 m, respectively, by using the tool of neighborhood analysis in ArcGIS 10.2. The ratio of the road area was calculated in each grid (Hoyo *et al*. 2011). All the selected human factors for each grid from year 2001 to 2016 were calculated using ArcGIS 10.2.

Vegetation coverage and land use factor

The Normalized Difference Vegetation Index (NDVI) was used to reflect the vegetation coverage of the study area. The NDVI data were derived from the MODIS NDVI product with a spatial resolution of 500 m provided by the Geospatial Data Cloud (http://www.gscloud.cn/). The land use data (1 km resolution) were obtained from the Resource and Environmental Data Cloud Platform (http://www.resdc.cn/Default.aspx), which provides the spatial distribution of vegetation types by digitizing the collections of vegetation type in China on the scale 1 : 1 million. Forest, including subtropical evergreen broad-leave forest type and mixed conifer and broad-leave forest type, covers about 64.95% of the total area. Vegetation type of shrub contains the subtropical evergreen broad-leave shrub, tropical evergreen broad-leaved shrub, and deciduous and broad-leaved shrub, taking about 20.40% of the study region. The cultivated land, fruit forest, and non-timber product forest are categorized into the cropland, covering about 12.6% of the total area. Grass (subtropical grass and tropical grass) only covers 1.69% of the total area. About 0.36% of the total area is the development land. According to the survey, regional forest coverage proportion is 66.80% and the agriculture land is about 11%, indicating the land classification of the raster data is close to reality. ArcGIS10.2 was used to calculate the proportion of each land cover in each grid.

2.2.3 Multicollinearity analysis among explanatory variables

We used the variance inflation factor (VIF) to detect the multicollinearity among the predictor variables before fitting the regression models. In general, a VIF above 10 indicates high correlations between explanatory variables (Guo *et al*. 2017). In this study, the socioeconomic variable GDP was removed because its VIF was 18.58, while other 17 predictor variables were used to fit both global and local beta regression models. The descriptive statistics of the 17 predictor variables were listed in Table 4.1. The spatial distributions of 17 predictor variables across the Fujian province are shown in Figure 4.3(b)-(r).

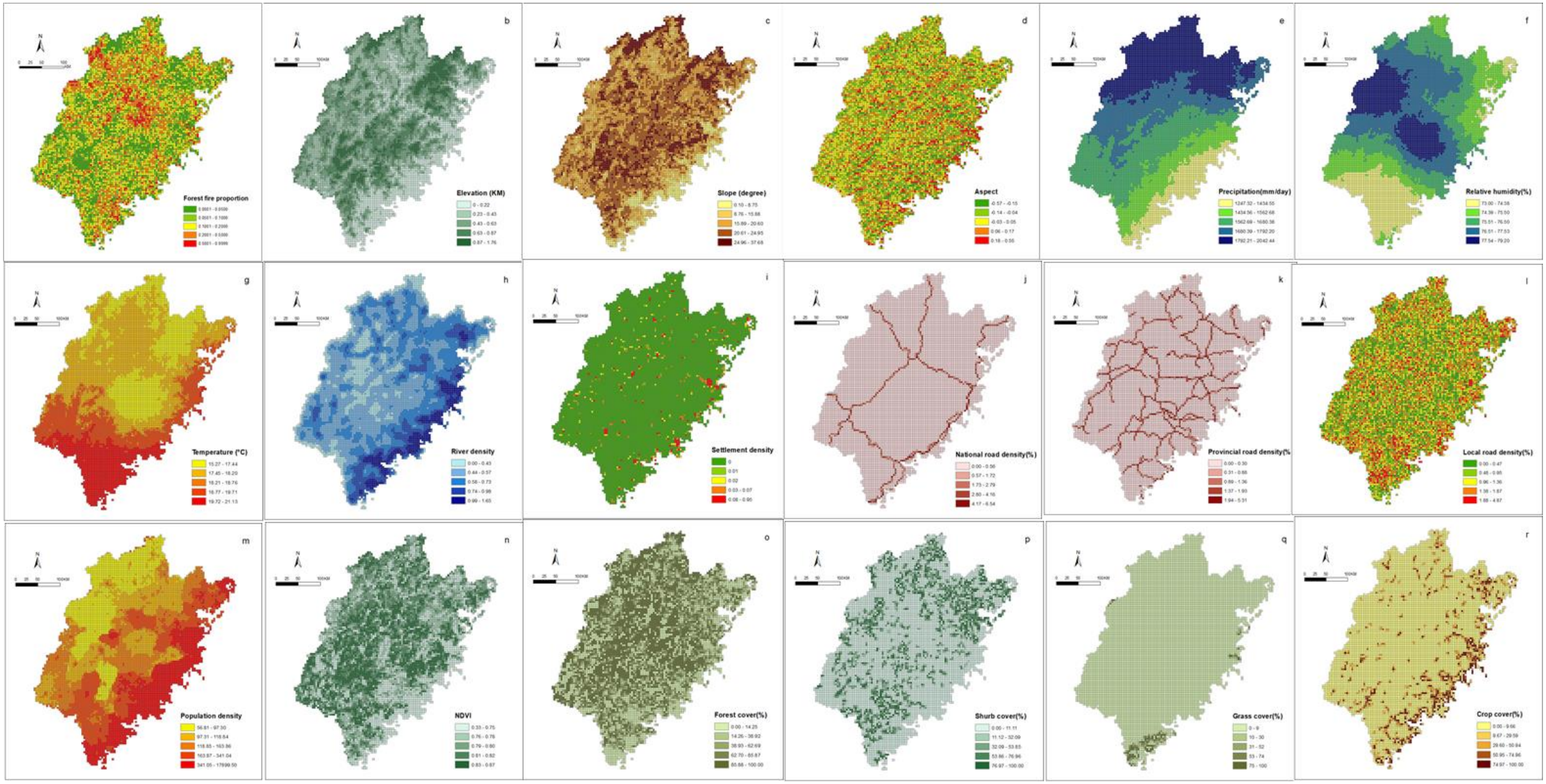


Figure 4.3 Spatial distributions of (a) Forest Fire Proportions, (b) Elevation, (c) Slope, (d) Aspect index, (e) Precipitation, (f) Relative humidity, (g) Temperature, (h) River density, (i) Settlement density, (j) National road density, (k) Provincial road density, (l) Local road density, (m) Population density, (n) NDVI, (o) Forest cover, (p) Shrub cover, (q) Grass cover, and (r) Crop cover.

3. Methodology

3.1 Theoretical background

3.1.1 Beta regression

Beta distribution is highly flexible and able to accommodate both unimodal and bimodal densities with varying degrees of skewness and heteroscedasticity. It is considered a natural choice for characterizing random variables within an interval (0, 1) such as proportion, percentage, or rate (Swearingen *et al*. 2011). Beta regression is a member of generalized linear models (McCullagh and Nelder 1989). Ferrari and Cribari-Neto (2004) proposed it for modeling continuous proportions based on the assumption that the response variable (*y*) follows a beta distribution. The beta distribution has the following probability density function (*pdf*):

$$f\left(y;p,q\right)=\frac{\Gamma\left(p+q\right)}{\Gamma\left(p\right)\Gamma\left(q\right)}y^{p-1}\left(1-y\right)^{q-1},\quad 0<y<1 \qquad [1]$$

where $p > 0$, $q > 0$ and $\Gamma(.)$ is the gamma function. The mean and variance of the beta random variable are:

$$E(y)=\frac{p}{(p+q)} \qquad [2]$$

$$Var\left(y\right)=\frac{pq}{\left(p+q\right)^{2}\left(p+q+1\right)} \qquad [3]$$

To model the mean of response variable more directly, Ferrari and Cribari-Neto (2004) re-parameterized the beta *pdf* by setting $\mu = p/(p+q)$ and $\phi = p+q$. Therefore, the Equations [2] and [3] are changed to:

$$E(y)=\mu \qquad [4]$$

$$Var\left(y\right)=\frac{\mu\left(1-\mu\right)}{1+\varphi} \qquad [5]$$

Thus, the beta *pdf* (Equation [1]) can be rewritten using the new parameters, e.g., μ represents the mean and ϕ is a precision parameter. If μ is fixed, a greater ϕ indicates a smaller variance of y (Ospina and Ferrari 2011). Then, the beta *pdf* with the new reparameterization is:

$$f(y;\mu,\varphi)=\frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma((1-\mu)\varphi)}y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1},\quad 0<y<1 \qquad [6]$$

where $0 < \mu < 1$ and $\phi > 0$. The response variable y is assumed within the interval (0, 1).

Let $y_1, \ldots, y_n$ be a random sample of the response variable which are independent from each other and each y_i, ($i = 1,\ldots, n$) follows the beta *pdf* (Equation [6]) with the mean μ and precision parameter ϕ. The mean of y_i can be linked to a linear function of predictor variables (X) such that:

$$g(\mu_i) = \eta = X\beta \qquad [7]$$

where β is a vector of unknown regression coefficients which can be estimated by maximum likelihood methods. Hence, $\eta = X\beta$ is a linear predictor and $g(\mu_i)$ is called a link function, which is a strictly monotonic and twice differentiable function that maps the interval of (0, 1). Several possible choices for the link function, in which the logit link function is commonly uses in practice,

$$g(\mu_i)=\ln\left(\frac{\mu_i}{1-\mu_i}\right) \qquad [8]$$

3.1.2 Geographically weighted regression (GWR)

To explore the spatial heterogeneity the data must have location coordinates (v_{xi}, v_{yi}) for each observation $\boldsymbol{i}$ ($\boldsymbol{i} = 1, 2, \ldots, n$). When the geographically weighted regression (GWR) was first developed, the Gaussian assumption was assumed for the model error term (Fotheringham *et al.* 1998), expressed as follows:

$$y_i=\beta_0(v_{xi},v_{yi})+\sum_{k=1}^{p}\beta_k(v_{xi},v_{yi})X_{ki}+\varepsilon_i \qquad [9]$$

where y_i is the response variable, X_k is a set of $\boldsymbol{p}$ predictor variables ($\boldsymbol{k} = 1, 2, \ldots, p$), $\beta_0(v_{xi}, v_{yi})$, $\beta_1(v_{xi}, v_{yi})$, …, $\beta_p(v_{xi}, v_{yi})$ are the regression coefficients for the $\boldsymbol{k}$th predictor variable at the $\boldsymbol{i}$th location, and ε_i

is the random error term whose distribution is N(0, σ^2I) with I denoting an identity matrix. To obtain the estimates of these functions for each predictor variable and each geographic location ***i***, the model fitting procedure of GWR is as follows: (1) draw a circle of a given radius around one particular location ***i*** (the center), (2) compute a weight (w_{ij}) for each neighboring observation ***j*** according to the distance d_{ij} between the location ***j*** and center ***i***, and (3) estimate the model coefficients using weighted least-square regression such that:

$$\hat{\beta}_i = \left(X'W_iX\right)^{-1} X'W_i y \qquad [10]$$

where the weight matrix W_i is:

$$W_i = \begin{pmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{pmatrix} \qquad [11]$$

The weighting function is defined by the kernel type and the size of kernel (bandwidth), which determines the geographical weight of the ***j***th neighboring observation at the ***i***th regression point. The weight should decrease gradually as the distance between ***i*** and ***j*** increases, until to a constant or zero. Parameter estimates are highly related to kernel size, so the choice of kernel is an important consideration. If W_i = I (i.e., identity matrix), each observation in the data has a weight of unity and the GWR model is equivalent to the ordinary least squares model. There are two common types of kernel function:

(i) Gaussian kernel with fixed bandwidth, in which each local regression model has the same spatial size of kernel, but each kernel may cover a different number of data points.

$$w_{ij} = e^{-\left(\frac{d_{ij}}{h}\right)^2} \qquad [12]$$

where d_{ij} is the distance between regression point ***i*** and neighbor ***j,*** and ***h*** is the bandwidth.

(ii) Adaptive methods with bi-square kernel, in which the bandwidth covers the same number of data points with non-zero weight within each regression model. This adaptive kernel is a common choice, especially when the sampling density varies greatly across space.

$$w_{ij} = \left[1-\left(\frac{d_{ij}}{h}\right)^2\right]^2, d_{ij} \leq h \quad \text{or} \quad w_{ij} = 0, \quad d_{ij} > h \qquad [13]$$

Essentially, GWR lets the data speak for themselves when estimating each regression coefficient β_{ik} for each geographic location ***i*** and each independent variable ***k***. Furthermore, the GWR methodology has been extended to the generalized linear modeling framework.

3.1.3 Geographically weighted beta regression (GWBR)

Silva *et al.* (2017) specified the GWBR as follows:

$$g(\mu_i) = \eta_i = \beta_0(v_{xi}, v_{yi}) + \sum_{k=1}^{p} \beta_k(v_{xi}, v_{yi}) X_{ki} + \varepsilon_i \qquad i = 1, \ldots, n \qquad [14]$$

where $g(.)$ is a link function that associated to the mean of response variable within the interval (0, 1), (v_{xi}, v_{yi}) represents the geographical coordinates of the ith observation, i = 1, …, n., and $\beta_k(u_i, v_i)$ is the local model coefficient for the kth predictor variable at the ith location. Using the local log-likelihood for the ith location (v_{xi}, v_{yi}), and the log-likelihood of beta regression (Ferrari and Cribari-Neto 2004),

$$L(\mu_i,\beta_k(v_{xi},v_{yi}),\varphi_i) = \log\Gamma(\varphi_i) - \log\Gamma(\mu_i\varphi_i) - \log\Gamma((1-\mu_i)\varphi_i)$$
$$+(\mu_i\varphi_i-1)\log y_i + ((1-\mu_i)\varphi_i-1)\log(1-y_i) \qquad [15]$$

where μ_i is the predicted mean at the location i.

Then, the local parameters can be estimated as the beta regression, using some nonlinear algorithm for optimization such as Newton or Quasi-Newton. More details on GWBR coefficient estimates and variances of regression coefficients are available in Silva *et al*. (2017).

3.2 Regression model

We choose the following regression model for both global beta and GWBR to investigate the relationships between forest fire proportions (y_i) and predictor variables:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{16} X_{16i} + \beta_{17} X_{17i} + \varepsilon_i \quad [16]$$

$$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i) X_{1i} + \ldots + \beta_{17}(u_i, v_i) X_{17i} + \varepsilon_i \quad [17]$$

The SAS procedure PROC GLIMMIX was used to fit the global beta and GWBR models (Equation [6]) (SAS Institute, Inc. 2014). The software GWR4.0, developed by Nakaya *et al.* (2014), was used to fit the GWR model.

3.3 Bandwidth selection for GWR and GWBR

In this study, we used Akaike Information Criterion (AIC) to determine the optimal bandwidth and related kernel function for estimating each regression coefficient for each geographic location *i* and each predictor variable (Fotheringham *et al*. 2002; Guo *et al*. 2008). A variogram model (Bailey and Gatrell 1995) was fitted in order to find the optimal bandwidth and kernel function. All optimal kernel functions and estimated bandwidths were chosen at the smallest AIC.

3.4 Assessment of spatial autocorrelation and nonstationary

Existence of spatial autocorrelation and heterogeneity can be evaluated from the model residuals of the global and local beta regression models using Moran's Index (Moran 1950). A positive Moran's I value indicates that a "high fire occurrence proportion (HOP)" pixel is neighboring with the HOP pixels, while a "low fire occurrence proportion (LOP)" pixel is neighboring with the LOP pixels. A negative Moran's I value indicates that a HOP pixel is neighboring with the LOP pixels, while a LOP pixel is neighboring with the HOP pixels.

To evaluate the spatial variation in the regression coefficients of GWBR, we followed the approach in Chen *et al*. (2012). The interquartile ranges (IQR) of the local model coefficients computed by GWBR were compared to the corresponding standard errors of the global beta model coefficients. When the IQR was twice as large as the standard error, it indicated that spatial non-stationarity existed in the relationship between the response variable and a specific predictor variable (Chen *et al*. 2012).

3.5 Model evaluation

Global beta and GWBR models were evaluated by: (1) mean squared of errors (MSE), (2) Akaike Information Criterion (AIC) (Sakamoto *et al*. 1986), (3) Pseudo R^2 for beta regression (Ferrari and Cribari-Neto 2004), and (4) correlation of generalized linear model (Zheng and Agresti 2000). The pseudo R^2 is the squared correlation of linear predictor and link-transformed response, specifically designed for beta regression (Ferrari and Cribari-Neto 2004). The correlation of generalized linear model (GLM) is a measure of predictive power for a GLM (Zheng and Argesti 2000). It follows:

$$\begin{aligned} \mathrm{Corr}(y, E(y \mid X)) &= \frac{\mathrm{Cov}(y, E(y \mid X))}{[\mathrm{Var}(y)\mathrm{Var}(E(y \mid X))]^{1/2}} \\ &= \left[1 - \frac{E[\mathrm{Var}(y \mid X)]}{\mathrm{Var}(y)}\right]^{1/2} \end{aligned} \quad [19]$$

where y is the observed value of the response variable, *E(y|X)* represents the conditional mean of y. The correlation between them equals to the positive square root of the average proportion of variance explained by the predictors.

Smaller values of MSE and AIC, greater values of pseudo R^2 and correlation indicate better model fitting and prediction.

4. Results

4.1 Global beta model

The estimated coefficients of all predictor variables for modeling the proportions of forest fire occurrence by the global beta model were listed in Table 4.2. All model coefficients of the topographical variables (i.e., elevation, slope, and aspect index) were negative, suggesting that the forest fire proportion would be decreased when any of them was increased. In contrast, the meteorology variables (i.e., precipitation, relative humidity, and temperature) had positive model coefficients, indicating the forest fire proportion would be increased if any of the meteorology variables was increased. For the human factors, only population density was negatively associated with the proportion of forest fire, meaning that the wildfire was more frequently happened in the region with low population density. Other human related factors, such as river density, settlement density, and road density (national, provincial, and local), played a positive role to the forest fire proportion. As any of them was increased, more frequent forest fires would be expected. The land coverage of forest, shrub, and grass were also positively related to the forest fire proportion, indicating that larger area of them would have higher chance of forest fire occurrence. On the contrary, the land cover of crop and NDVI were negatively associated to the forest fire proportion, implying that if the area had more vegetation or crop, the proportion of forest fire would go down (Table 4.2).

In terms of statistical significance, the important predictor variables were elevation, slope, aspect index, precipitation, relative humidity, river density, settlement density, national road density, population, NDVI, grass cover, and crop cover. On the other hand, the temperature, provincial and local roads, and land coverage of forest and shrub were not statistically significant to the forest fire proportion (Table 4.2).

Table 4.2 Estimated coefficients of global beta regression.

Variable	Estimate	Standard Error	*p*-value
Intercept	-9.9264	1.5208	<.0001
Elevation	-1.5264	0.06736	<.0001
Slope	-0.0118	0.00414	0.0043
Aspect Index	-0.2834	0.1009	0.005
Precipitation	0.00089	0.00012	<.0001
Humidity	0.1035	0.01391	<.0001
Temperature	0.01761	0.02281	0.4401
River Density	0.1665	0.07974	0.0368
Settlement Density	0.02645	0.00392	<.0001
National Road Density	0.0424	0.01728	0.0142
Provincial Road Density	0.04465	0.02479	0.0717
Local Road Density	0.00762	0.02161	0.7244
Population	-0.4501	0.04435	<.0001
NDVI	-0.7397	0.3457	0.0324
Forest Cover	0.02754	0.036	0.4443
Shrub Cover	0.02498	0.04824	0.6046
Grass Cover	0.4022	0.1351	0.0029
Crop Cover	-0.3925	0.06714	<.0001
Scale	3.2651	0.05854	

4.2 GWR and GWBR models

The summaries of estimated coefficients of both GWR and GWBR models were displayed in Table 4.3. The coefficient estimates of the predictor variables in GWR generally showed a similar trend or pattern to the global beta model in terms of the mean or median of the model coefficients, except temperature, local road density, forest cover, and shrub cover density. While using the GWBR model, the mean and median of estimated coefficients of relative humidity, temperature, local road density, forest cover, and shrub cover density showed opposite directional trends compared to the global beta model.

Because both GWR and GWBR models produced the model coefficients for each location (pixel) across the study area, the spatial variation of the local model coefficients can be evaluated by comparing their IQRs and the standard errors of the global model coefficients. For the GWBR model, the IQRs of the local coefficients of all predictors were at least twice larger than the corresponding standard errors of the global beta model coefficients, indicating that all topographical, meteorological, human related, vegetable, and land coverage factors were spatially heterogeneous across Fujian Province. The estimated model coefficients from GWR were also spatially varied across the region, except the local road density (Table 4.3).

We constructed the spatial maps of the local model coefficients which had the local t-test exceeded ±1.96 (i.e., statistically significant) for each predictor variable for GWBR (Figure 4.4) and GWR (Figure 4.5). For the GWBR models, the forest fire proportions were significantly negatively related to the elevation almost across the whole regions of Fujian province, except some flat terrain along the southeast coast, where the big cities are located. Moreover, that relationship became stronger from southeast to northwest. The slope is essentially a negative factor in north and west, positive on the southern corner, but not important in east. The aspect index did not show a clear pattern across the study area as the elevation and slope did, except few significant clusters in southwest negatively associated with the forest fire proportion. All three meteorological predictor variables were significantly related to the forest fire proportions across the majority of Fujian Province, with different spatial magnitude patterns. Generally, the precipitation played a positive role in west, but negative in east. The relative humidity positively impacted on the forest fire proportion in north, while negatively affected it in south. The temperature was negatively associated in the center area, but changed to positive in northwest, northeast, and south corner. For the human factors, the population showed the different relationships with the forest fire proportion between northwest and southeast. The river density had negative association in the central study area, but gradually changed to positive toward to the edge of Fujian Province. Other human explanatory variables (i.e., settlement density, national, provincial, and local road density) presented significances in some areas, but did not reveal clear pattern or trend. Vegetation coverage (NDVI) was positively related to the forest fire proportion from southwest to northeast, negatively related in northwest. For the four land cover types, forest, shrub, and crop had significant influences on more than half of the study area, but grass only had few positive spots in south and northwest. Significant forest cover was scattered across the study area. Shrub cover was positively linked in the majority of south and some spots in north, while negative linked in the central areas. In general, connection between crop cover and the forest fire proportion changed from positive to negative from northwest to southeast.

For the GWR models, the elevation was statistically significant in northwest, and NDVI was an important factor in southeast. Though the other factors had some local important relationships with the forest fire proportions, they were scattered across the study region without a clear trend or pattern.

Table 4.3 Summary statistics for estimated local coefficients from GWR and GWBR models and relative spatial variation status.

Statistic	Model	$\beta_{\text{Elevation}}$	β_{Slope}	$\beta_{\text{Aspect index}}$	$\beta_{\text{Precipitation}}$	$\beta_{\text{Relative Humidity}}$	$\beta_{\text{Temperature}}$	$\beta_{\text{River density}}$	$\beta_{\text{Settlement density}}$	$\beta_{\text{National road density}}$	$\beta_{\text{Provincial road density}}$	$\beta_{\text{Local road density}}$	$\beta_{\text{Population}}$	β_{NDVI}	$\beta_{\text{Forest cover}}$	$\beta_{\text{Shrub cover}}$	$\beta_{\text{Grass cover}}$	$\beta_{\text{Crop cover}}$
Mean	GWR	-0.217	-0.0007	-0.0221	0.00012	0.00669	-0.0134	0.02828	0.00523	0.007147	0.00157	-0.00242	0. 035	-0.1944	-0.0051	-0.0102	0.0486	-0.0152
	GWBR	-1.705	-0.0056	-0.2495	0.00137	-0.0788	-0.1006	0.2261	0.03938	0.01744	0.01337	-0.02232	-0.0773	-1.429	-0.0123	-0.0356	7.082	-0.2448
Median	GWR	-0.211	-0.0009	-0.0214	0.00017	0.00371	-0.0057	0.02279	0.00372	0.00683	0.00303	-0.00276	-0.016	-0.2498	-0.0028	-0.0102	0.0419	-0.0181
	GWBR	-1.659	-0.0066	-0.2457	0.00082	-0.0544	-0.1450	0.3163	0.02687	0.02807	0.01853	-0.02264	-0.2877	-1.970	-0.0151	-0.0227	0.105	-0.2707
Min	GWR	-0.395	-0.0113	-0.0879	-0.0012	-0.0489	-0.0897	-0.1039	-0.0033	-0.01283	-0.02902	-0.01476	-0.504	-1.0786	-0.0374	-0.0698	-1.3325	-0.1091
	GWBR	-6.107	-0.1774	-3.4203	-0.0362	-1.6507	-1.5989	-4.9711	-0.5566	-1.7668	-0.7378	-0.6158	-17.477	-14.423	-0.8317	-1.6921	-234.07	-2.2976
Max	GWR	0.097	0.00899	0.07520	0.00142	0.05459	0.1187	0.16134	0.03135	0.03238	0.02636	0.01104	0.638	0.5804	0.0368	0.0796	0.3027	0.1653
	GWBR	4.640	0.1719	2.0256	0.0354	1.1763	3.4335	3.4992	1.7956	3.9569	0.6567	0.5231	28.539	19.026	1.1086	1.2803	20220	1.9970
IQR	GWR	0.126	0.00515	0.03627	0.00041	0.03066	0.0376	0.07608	0.00637	0.007626	0.008226	0.00566	0.130	0.51104	0.0133	0.035	0.3533	0.0404
	GWBR	1.7824	0.0736	0.7634	0.00588	0.5900	0.7571	1.3701	0.05283	0.14496	0.21072	0.17265	0.3074	6.1249	0.2918	0.5101	1.1602	0.6092
Ste†·	GWR	0.00903	0.00058	0.01354	1.60E-05	0.0019	0.00306	0.01098	0.00053	0.00235	0.00337	0.00294	0.0045	0.04695	0.0048	0.0065	0.0178	0.009
	GWBR	0.06736	0.00414	0.1009	0.00012	0.01391	0.02281	0.07974	0.00392	0.01728	0.02479	0.02161	0.0444	0.3457	0.036	0.0482	0.1351	0.06714
Status	GWR	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	S	NS	NS	NS	NS	NS	NS
	GWBR	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS

† Note: Standard error (Ste) was estimated from the global model.

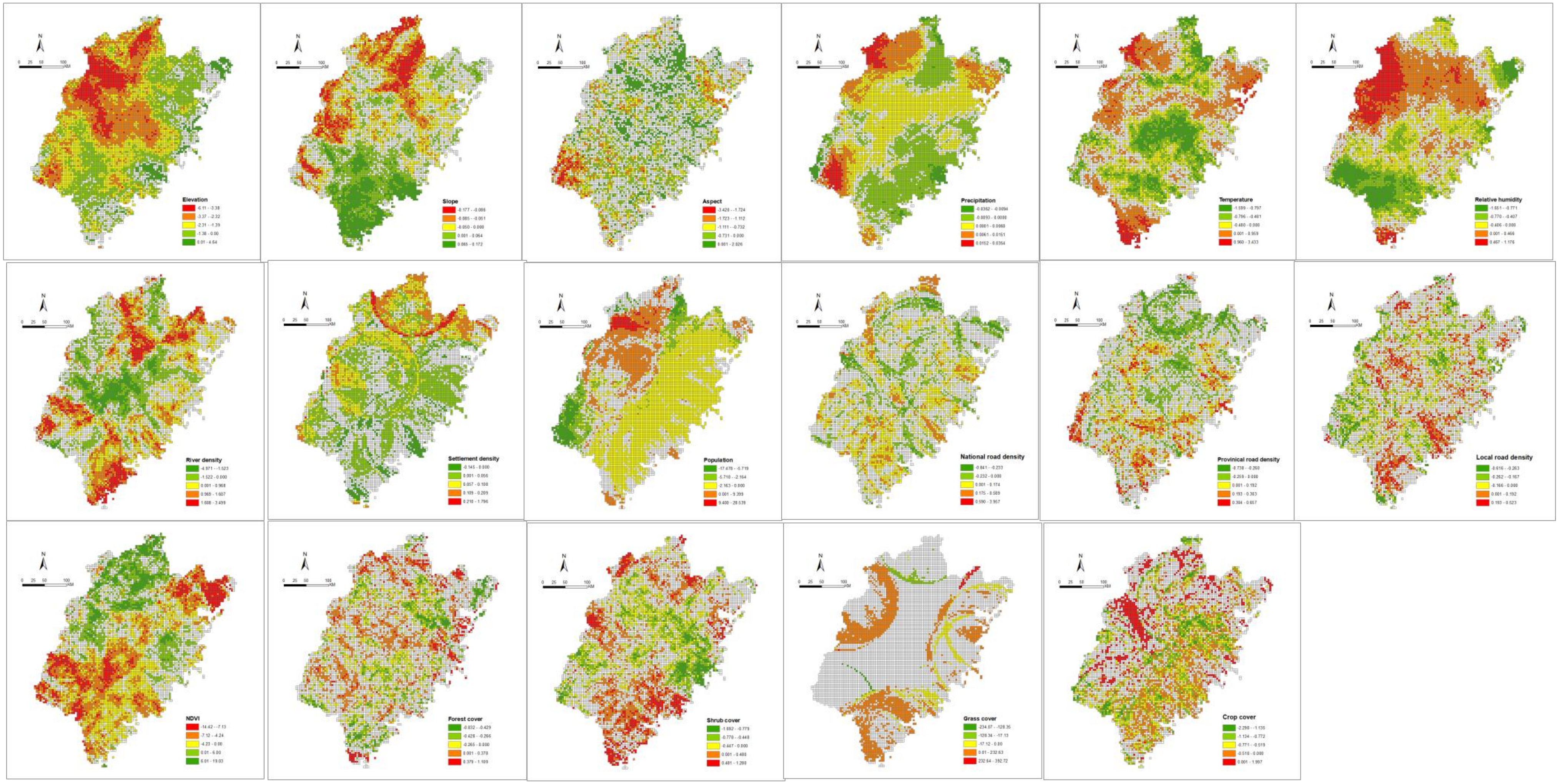


Figure 4.4 Geographical maps for significant coefficients (±1.96) of predictors based on GWBR model.

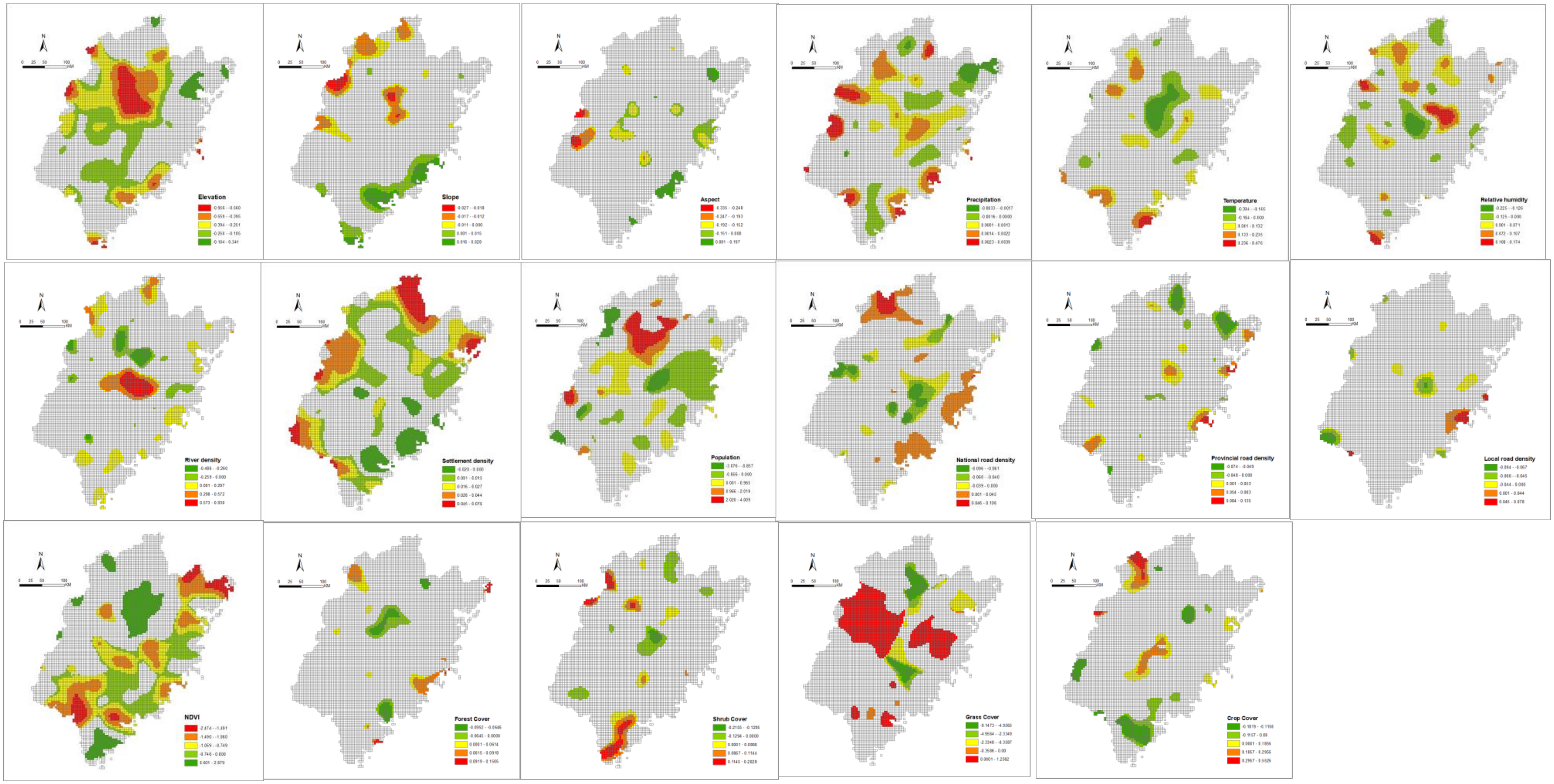


Figure 4.5 Geographical maps for significant coefficients (±1.96) of predictors based on GWR model.

4.3 Model fitting and predictive performance

The performance of model fitting was listed in Table 4.4. Smaller MSE or/and AIC indicates better model fitting. However, the model fitting performance was mixed among the three models, depending on which statistics were used. The MSE of global beta, GWBR, and GWR were 0.0188, 0.0282, and 0.0103 respectively. The AIC of global beta, GWBR, and GWR models were -19844.66, -19734.34, and -10676.07, separately. Therefore, GWR fitted the data better than both global beta and GWBR in terms of MSE, but GWR had the largest AIC (i.e. worst model fitting) among the three models. Rather, the global beta had smaller AIC than both GWBR and GWR (Table 4.4).

Pseudo R^2 and the correlation of generalized linear model were used to compare the model fitting and predictive power between global beta and GWBR. The pseudo R^2 of the global beta was 12.63%, and GWBR was 11.56%. The correlations of the global beta and GWBR were 39.24% and 31.23%, respectively. The global beta model had larger pseudo R^2 and correlation, indicating that it performed better in model fitting and prediction than GWBR (Table 4.4).

Table 4.4 Model fitting performance of Global beta, GWBR, and GWR.

Statistics	**Global beta**	**GWBR h = 100,000 m**	**GWR h = 16,464.29 m**
MSE	0.0188	0.0282	0.0103
AIC	-19844.66	-19734.34	-10676.07
Pseudo R^2	0.1263	0.1156	0.5385 (R^2)
Correlation	0.3924	0.3123	0.7376

Figure 4.6 illustrates the frequency distributions of the observed forest fire proportion and predicted forest fire proportion from global beta, GWR, and GWBR models. It revealed that the GWR model generated some predicted values smaller than 0, outside the interval of forest fire proportion (0, 1). On the other hand, both global beta and GWBR models predicted forest fire proportions within (0, 1). With regards to the shape of frequency distribution, the predictions based on the global beta and

GWBR models were skewed to right, similar to the observed forest fire proportion, while the predictions from GWR were less skewed or relatively normal distributed.

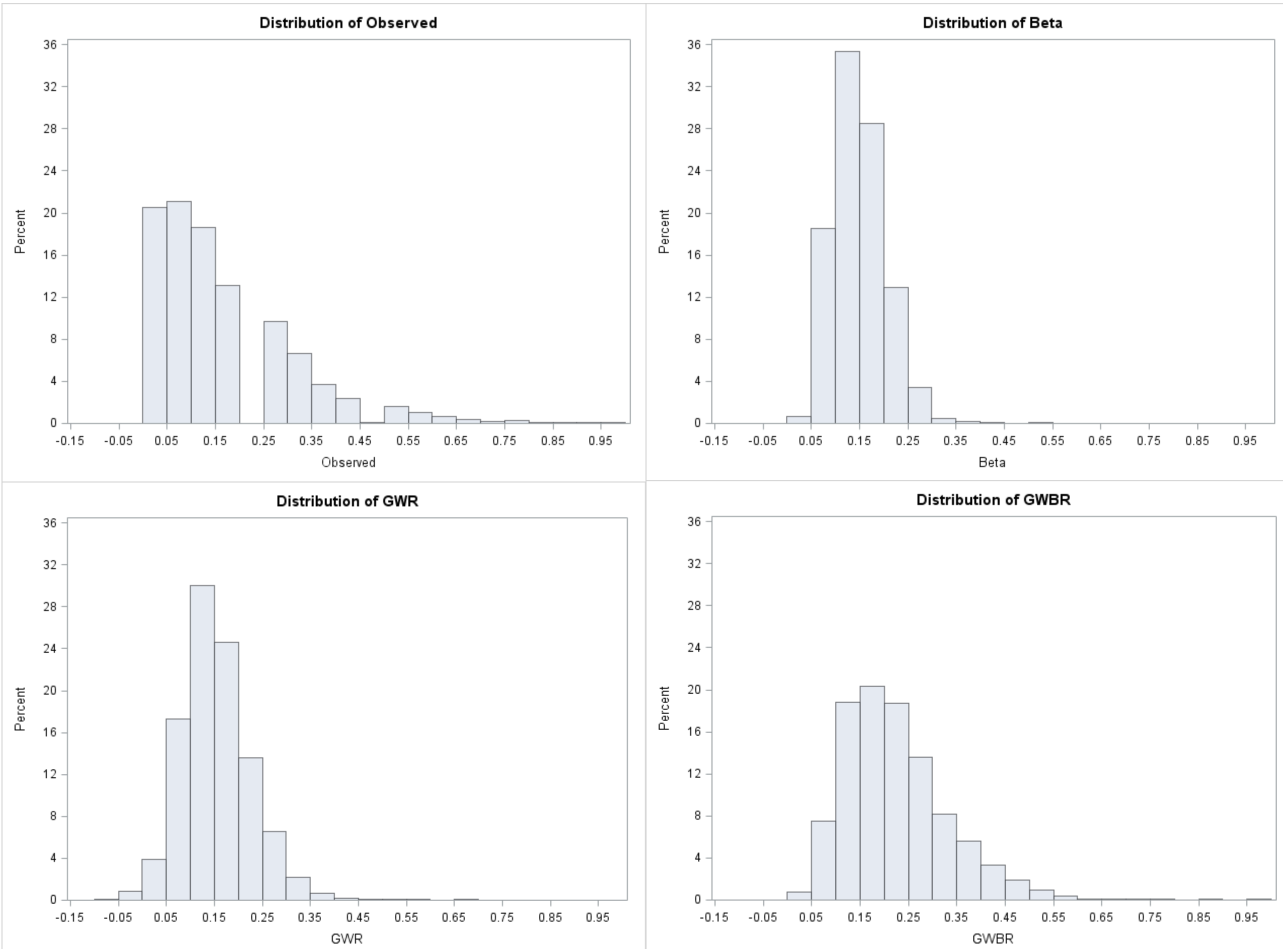


Figure 4.6 Frequency distributions of observed and predicted forest fire proportions from global beta, GWR, and GWBR models.

5. Discussion

5.1 Model comparison for proportion of forest fire occurrence

Our results implied that the global beta and GWR models were relatively better in model fitting and prediction than GWBR models for the forest fire proportion in this study, but both of them have limitations. For the global beta model, it derived the average relationships between response variable (i.e., proportion of forest fire occurrence) and environmental factors across the whole study region and assumed the relationships invariant over space. However, the existence of spatial non-stationary of the response variable (Moran's I = 0.044, $p < 0.0001$) and variations of predictor variables suggested the demand of local statistics and models to detect the localized clustering around an individual location, especially in a large area when the single measure of global association may contribute little meaning. The spatial heterogeneity of the model coefficients from both GWR and GWBR models (Table 4.3) confirmed the necessary and appropriateness to use local models to investigate the spatially varying relationships between the proportion of forest fire and predictors.

The GWR and GWBR models provided local significance tests on the model coefficients, which would help us to target the geographical spots where a given predictor may have different impacts on the response variable. For instance, the temperature was not an important factor based on the global beta model, but was statistically significant within some regions, e.g., the temperature negatively impacted the forest fire proportion in the central area of the Fujian Province, while positively impacted it in northwest, northeast, and south corner. Therefore, the GWR and GWBR models were preferred to detect the local association. Those significant local relationships could also present some pattern or trend connected with the study area. For example, the strength of elevation significantly related to the forest fire proportion based on the GWBR models was decreased from northwest to southeast. That pattern was reasonable according to the terrain of Fujian Province, where the mountains are clustered in northwest and cities are located along southeast. However, the GWR models did not produce the relationship patterns between forest fire proportions and environmental factors as clear as the GWBR models. Instead, most of the geographical maps of significant predictors of the GWR models are scattered and sparse, resulting the difficulties for interpretation. Thus, the GWBR model performed more appropriately in targeting the geographical spots where the environment factors are significant and important to the proportion of forest fire occurrence.

It is not surprised that GWR had the smallest MSE, highest R^2, and correlation (Table 4.4), since the localized ordinary least square regression is designed by minimizing the sum of squared errors. However, the data of forest fire proportions not only is restricted within the unit interval (0, 1), but also commonly have varying degrees of skewness and heteroscedasticity (Figure 4.2). Beta regression is a relatively new and modern regression method designed for proportion or rate response variables (Ferrari and Cribari-Neto 2004). The advantages of beta regression include that it is naturally heteroskedastic, can easily accommodate the varying degrees of skewness, and is highly flexible to model both unimodal and bimodal distributions of proportion data. The smaller AIC values of global beta and GWBR models proved that they were superior to GWR for modeling the forest fire proportions. In addition to model fitting, GWR had some negative predictions, which exceeds the lower limit zero, violating the definition of proportion between 0 and 1. In addition, the frequency distribution of GWR predictions was relatively normal, which did not match the observed forest fire proportion skewed to right. In summary, the large AIC, negative predicted value, normal distributed frequency of model predictions, and difficult interpretation of significant local relationship implied that GWR was not appropriate to model the forest fire proportions.

5.2 Influence of drivers on proportion of forest fire in Fujian

In this study the response variable, proportion of forest fire occurrence, was used evaluate the frequent levels of wildfire occurrence within a time interval (from 2001 to 2016) in Fujian province.

Both global beta and GWBR models were applied to investigate the environment drivers and their influences on the proportions of fire occurrence.

According to the global beta model, the proportion of the forest fire occurrence was increased with lower topographical factors (i.e., elevation, slope, and aspect index), greater meteorological factors (i.e., precipitation, humidity, and relative humidity), more human manufacturing factors (i.e., river, settlement, national, provincial and local road density), smaller population, less vegetation coverage and cropland coverage, but more forest, shrub and grassland. However, the GWBR models showed that they were not constantly positive or negative across whole study area. For example, the significant model coefficients of population revealed a trend that the forest fire proportion was highly likely to happen with more people in northwest, but less chance with more people in southeast. One reason is that the developed cities are clustered in southeast. Denser population would cause reduced forest cover and result in smaller possibility of forest fires. While in northwest where most regions are covered by forests, the growth of population would create more human activities so increasing the chance of human-caused forest fires. The signs of the statistically significant model coefficients of three meteorological factors also changed across the study area. The precipitation and relative humidity had positive effects in northwest, but negative in southeast. One explanation is that the relative humidity affects the growth of ground cover vegetation. Higher relative humidity is beneficial to the growth of ground cover vegetation, which increases the fuel load (Su *et al.* 2019). Since the majority of the northwest are forests, the ground is covered with vegetation. More rainfall and humidity fostered the surface fuel load, increasing the probability of forest fires. In contrast, the developed cities were along southeast, where the ground vegetation is too small to produce the surface fuel load so that more precipitation and humidity would lessen the chance of forest fires. In addition to the varied signs of the influential factors, GWBR also provided significant local factors which were not revealed by the global beta model. Specifically, the temperature, forest cover, and shrub cover had essential roles to the forest fire probability in most regions. But the grass cover, an important factor estimated by the global beta model, only regionally significant in a part of southeast, where have more grasslands.

5.3 Shortcomings

There were other limitations in this study. Firstly, the beta regression is designed to model all proportions or rates between (0, 1), but neither 0 nor 1 is included. It means that the value of response variable modeled by the beta regression can be close to 0 and 1, say .001 or .998, but not 0 or 1 exactly. In our study, there are a total of 1525 locations with no fire occurrence and 2 regions had 100% proportions of fore fire occurrence. To keep all proportion of forest fire occurrence within the interval of (0, 1), we had to change 0 into 0.0001 and 1 into 0.9999. Thus, about 20% of raw data were changed. Although the percentage of the changed data was relatively small, the impact to model fitting and prediction was unknown. A version of beta regression model, called Zero-One-Inflated beta regression (Ospina and Ferrari 2012; Swearingen *et al*. 2012), may work in this situation. It's one of those models that has been around in theory so far, but is only in the past few years the technique became available in (some) mainstream statistical software. It is a type of mixture model that has three modeling processes. One is a process that distinguishes between zeros and larger than zeros. Likewise, there is a process that distinguishes between ones and less than one. And then there is a third process that determines between zero and one (not including zero and one). The first and second processes can be modeled via a logistic regression and the third is a beta regression. These three models should be fitted run simultaneously. They can each have their own set of predictors and their own set of predictors. Zero-One-Inflated beta model would be a good alternative to globally model the forest fire proportions. However, it would be difficult to use it to explore the spatially varying relationships under a GWR framework. Following the process of Zero-One-Inflated beta model, three bandwidths should be chosen for GWR Zero-Inflated Poisson, GWR One-Inflated Poisson, and GWR beta models, respectively. The raw data is divided into three parts: all zeros, all ones, and all between zero and one.

The split of response variable makes the difficult of model comparison to global beta and GWR models. Additionally, the bandwidth selections for GWR Zero-Inflated Poisson, GWR One-Inflated Poisson, and GWR beta models also will be a challenge.

The second main concern is the bandwidth selection. In this study, we used AIC value as the criteria for the GWR model, resulted to the bandwidth 16,464 m to define how many pixels nearby were included to compute the local model to the focal location. But when we applied that bandwidth to GWBR model, it cannot be estimated from beta log-likelihood. Silva and Lima (2017) developed a SAS marco of GWBR, which contains the bandwidth search function. However, it cannot work for our data (a total of 7433 observations and 17 explanatory variables) due to the limitation of SAS memory. Therefore, we used the spatial variogram to find the reasonable possible bandwidth (Figure 4.7), which was between 100,000 m and 200,000 m using the spherical kernel function. We subjectively selected 20 bandwidths between them, and found that bandwidth of 100,000 m yielded best results, closer to the global beta model. However, 100,000 m might not be the best choice if all possible bandwidths were simulated. This can be a good research topic in the future.

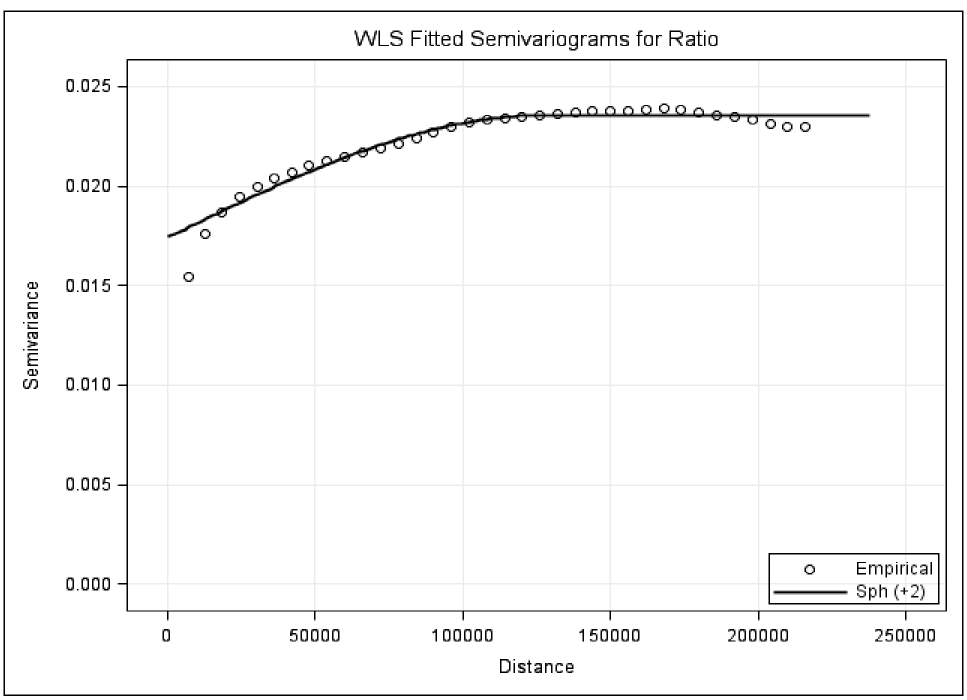


Figure 4.7 Variogram and fitted spherical kernel line for observed probability of forest fire occurrence.

As mentioned before, most of statistic models for evaluating the probability of forest fire were logistic or modified logit model, this study may be the first application of beta regression and geographically weighted beta regression. Because the logit model accentuated on probability while beta regression was quantitative value, it was impossible to compare these two statistic tools. Therefore, there was not enough reference to assess the beta regression application in forest fire proportion.

6. Conclusion

This study investigated both global and local relationship between the proportion of forest fire occurrence and relevant topographical, meteorological, human, vegetation coverage, and land coverage factors by the global beta, GWR and GWBR models. The response variable, the proportion of forest fire, is designed to evaluate the probability of wildfire occurrence during the time period from year 2001 to 2016 in Fujian province, China. The results indicated that global beta and GWBR models displayed better model fitting and prediction performances than the classical GWR model for the rate / proportion response variable. In addition to model performance, GWBR was well at targeting the essential hotspots of predictor variables.

The drivers and its spatially varied association to the proportion of forest fire occurrence were also explored by the GWBR model. Generally, the likelihood of wildfire is higher in lower elevation, stronger sunshine (meaning smaller aspect index), denser settlement, and less cropland coverage. Other factors had different influence on probability of forest fire through the study region. In summary, GWBR is an appropriate method to analyze the proportion of forest fire occurrence which spatially varied and within range between (0, 1).

7. References

Bailey T.C., Gatrell A.C., 1995. Interactive spatial data analysis. Longman Scientific and Technical. Routledge.

Bayes, C., Bazán J.L., García C., 2012. A new robust regression model for proportions. *Bayesian Analysis* 7(2): 771-796.

Branscum, A. J., Johnson, O. J., Thurmond, M., 2007.Bayesian beta regression: applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics* 49: 287-301.

Brunsdon C., Fotheringham A.S., Charlton M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* 28(4): 281-298.

Brunsdon C., Fotheringham A.S., Charlton M.E., 1998. Geographically weighted regression - Modelling Spatial Non-Stationarity. *The Statistician* 47(3): 431-443.

Calkin, D.E., Cohen J.D., Finney M.A., Thompson M.P., 2014. How risk management can prevent future wildfire disasters in the wildland-urban interface. *Proceedings of the National Academy of Sciences of the United States of American* 111(2): 746-751.

Catry F.X., Rego F.C., Bação F.L., Moreira F., 2009. Modeling and mapping wildfire ignition risk in Portugal. *International Journal of Wildland Fire* 18: 921-931.

Chang Y., He H.S., Bishop I., Hu Y., Bu R., Xu C., Li X., 2007. Long-term forest landscape responses to fire exclusion in the Great Xing'an Mountains, China. *International Journal of Wildland Fire* 16: 34-44.

Chang, Y., Zhu, Z., Bu, R. Chen H., Feng Y., Li Y., Hu Y., Wang Z., 2013. Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landscape Ecology* 28: 1989-2004.

Chen V.Y., Deng W.S., Yang T.C., Matthews S.A., 2012. Geographically weighted quantile regression (GWQR): An application to mortality data. *Geographical Analysis* 44(2): 134-150.

Guo F., Innes L.J., Wang G., Ma X., Sun L., Hu H., Su Z., 2015. Historic distribution and driving factors of human-caused fires in the Chinese boreal forest between 1972 and 2005. *Journal of Plant Ecology* 8: 480-490.

Guo F., Su Z., Wang G., Sun L., Lin F., Liu A., 2016. Wildfire ignition in the forests of southeast China: Identifying drivers and spatial distribution to predict wildfire likelihood. *Applied Geography* 66: 12-21.

Guo F., Su Z., Wang G., Sun L., Tigabu M., Yang X., Hu H., 2017. Understanding fire drivers and relative impacts in different Chinese forest ecosystems. *The Science of the Total Environment* 605-606: 411-425.

Guo F., Ju Y., Wang G., Alvarado E.C., Yang X., Ma Y., Liu A., 2018. Inorganic chemical composition of PM2.5 emissions from the combustion of six main tree species in subtropical China. *Atmospheric Environment* 189: 107-115

Ferrari, S.L.P, Cribari-Neto, F., 2004. Beta Regression for modelling rates and proportions. *Journal of Applied Statistics* 31(7): 799-815.

Ferrari, S.L.P, 2013. Beta regression modeling: recent advances in theory and applications. Available at https://www.ime.usp.br/~sferrari/13EMRslidesSilvia.pdf.

Fotheringham A.S., Charlton M.E., Brunsdon C., 1996. The geography of parameter space: an investigation of spatial non-stationarity. *International journal of Geographical Information System* 10: 605-627.

Fotheringham A.S., Brunsdon C., Charlton M., 2002. Geographically weighted regression. Wiley, New York. 284 p.

GWR4.0 2014. Application for Geographically Weighted Regression Modelling. Version 4.0.80. Developed at National Centre for Geocomputation, National University of Ireland Maynooth and Department of Geography, Ritsumeikan University, Japan.

Hoyo V.L., Martín Isabel M., Martínez Vega F., 2011. Logistic regression models for human-caused wildfire risk estimation: analysing the effect of the spatial accuracy in fire occurrence data. *European Journal of Forest Research* 130: 983-996.

Hutchinson M.F., 2004. ANUSPLIN Version 4.3. Centre for Resource and Environmental Studies, Australian National University, 23 November 2005. Available at https://fennerschool.anu.edu.au/research/products/anusplin

Lozano F.J., Suárez-Seoane S., de Luis E., 2007. Assessment of several spectral indices derived from multi-temporal landsat data for fire occurrence probability modelling. *Remote Sensing of Environment* 107: 533-544.

Martell D.L., Otukol S., Stocks B.J., 1987. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. *Canadian Journal of Forest Research* 17: 394-401.

McCoy V.M., Burn C.R., 2005. Potential alteration by climate change for forest-fire regime in the boreal forest of central Yukon Territory. *Arctic* 58: 276-285.

McCullagh P., Nelder J.A., 1989. Generalized Linear Models. Second Edition. Chapman and Hall.

McKenney D.W., Pedlar J.H., Papadopol P., Hutchinson M.F., 2006. The development of 1901-2000 historical monthly climate models for Canada and the United States. *Agricultural and Forest Meteorology* 138: 69-81.

Moran, P., 1950. A Test for the Serial Independence of Residuals. Biometrika 37: 178-181.

Ospina, R., Ferrari S.L.P, 2012. A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis* 56: 1609-1623.

Pastor E., Zarate L., Planas E., Arnaldos J., 2003. Mathematical models and calculation systems for the study of wildland fire behavior. *Progress in Energy and Combustion Science* 29: 139-153.

Podur J., Martell D.L., Csillag F., 2003. Spatial patterns of lightning-caused forest fires in Ontario, 1976-1998. *Ecological Modelling* 164: 1-20.

Preisler H.K., Brillinger D.R., Burgan R.E., Benoit J.W., 2004. Probability based models for estimation of wildfire risk. *International Journal of Wildland Fire* 13(2): 133-142.

Preisler H.K., Westerling A.L., 2007. Statistical model for forecasting monthly large wildfire events in western United States. *Journal of Applied Meteorology and Climatology* 46: 1020-1030.

Preisler H.K., Westerling A.L., Gebert K.M., Munoz-Arriola F., Holmes T.P., 2011. Spatially explicit forecasts of large wildland fire probability and suppression costs for California. *International Journal of Wildland Fire* 20: 508-517.

Pyne, S.J., Andrews, P.L., and Laven, R.D., 1996. Introduction to Wildland Fire. John Wiley & Sons, Inc. New York. 769 p.

Ríos-Pena L., Kenib T., Cadarso-Suárez C., Klein N., Marey-Pérez M., 2018. Studying the occurrence and burnt area of wildfires using zero-one-inflated structured additive beta regression. *Environmental Modelling and Software* 110: 107-118.

Rydlewski, J.P., 2007. Beta-regression model for periodic data with a trend. *Universitatis Iagellonicae Acta Mathematica* 1298(45): 207-217.

Sakamoto, Y., Ishiguro, M., Kitagawa, G. 1986. Akaike information criterion statistics. D. Reidel Publishing Company, Dordrecht, Netherlands.

SAS INSTITUTE, INC. 2014. SAS/STAT users' guide. Version 9.1.3. SAS Institute, Inc. Cary, NC.

Scholze M., Knorr W., Arnell N.W., Prentice I.C., 2006. A climate-change risk analysis for world ecosystems. *Proceedings of the National Academy of Science of the United States of America* 103: 13116-13120.

Silva A.R., Lima A.O., 2017. Geographically weighted beta regression. *Spatial Statistics* 21: 279-303.

Souza, D.F., Moura, F.A.S., 2012. Multivariate beta regression. Technical Report. Available at http://www.dme.im.ufrj.br/arquivos/publicacoes/arquivo245.pdf.

Su Z., Hu H., Tigabu M., Wang G., Zeng A., Guo F., 2019. Geographically weighted negative binomial regression model predicts wildfire occurrence in the Great Xing'an mountains better than negative binomial model. *Forests* 10: 377-394.

Swearingen C.J., Tilley B.C., Adams R.J., Rumboldt Z., Nicholas J.S., Bandyopadhyaya D., Woolson R.F., 2011. Application of beta regression to analyze Ischemic stroke volume in NINDS rt-PA clinical trials. *Neuroepidemiology* 37(2): 73-82.

Vega Garcia C., Woodard P.M., Titus S.J., Adamowicz W.L., Lee B.S., 1995. A logit model for predicting the daily occurrence of human caused forest fires. *International Journal of Wildland Fire* 5(2): 101-111.

Westerling A.L., Hidalgo H.G., Cayan D.R., Swetnam T.W. 2006. Warming and Earlier Spring Increase Western U.S. Forest Wildfire Activity. *Science* 313(5789): 940-943.

Xystrakis F., Koutsias N., 2013. Differences of fire activity and their underlying factors among vegetation formations in Greece. *iForest-Biogeosciences and Forestry* 6: 132-140.

Zhang X., Kang S., Zhang L., Liu J., 2010. Spatial variation of climatology monthly crop reference evapotranspiration and sensitivity coefficients in Shiyang river basin of northwest China. *Agricultural Water Management* 97: 1506-1516.

Zimprich, D., 2010. Modeling change in skewed variables using mixed beta regression models. *Research in Human Development* 7: 9-26.

Zhang, B., Agresti A., 2000. Summarizing the predictive power of a generalized linear model. *Statistics in Medicine* 19: 1771-1781.

Chapter V: Summary and Conclusion

This dissertation presents the applications of geographically weighted generalized linear models and geographically weighted quantile regression for modeling the occurrence of forest fires in Fujian province, China. The first study was on the applications and comparisons of global Poisson and negative binomial models against geographically weighted Poisson (GWPR), and geographically weighted negative binomial (GWNBR) models to determine the spatially varying relationships between the counts of forest fire and topographical, meteorological, human, vegetation coverage, and land cover factors. The results implied that, in general, more forest fires occurred with lower elevation, flatter terrain, and higher population density areas. Across the study area, the count of forest fire had the problems of overdispersion and spatial non-stationarity. The assessment of model fitting and predictions showed that GWNBR fitted the dispersed forest fire count data better than other models, produced more precise and stable model parameter estimation, and yielded more realistic spatial distributions of model predictions.

Generalized linear models evaluated the relationships between forest fire counts at a mean or average level and related driving factors. However, people are also interested in learning the risks of forest fire at different quantile levels in order to gain a full knowledge of possible risk levels of the forest fire occurrence, which can be achieved by quantile regression. In the second study, we applied the global and geographically weighted quantile regression (GWQR) models to investigate the spatially varying relationships at the 50th, 75th, 90th, and 99th quantile levels of forest fire risks. Our results showed that even the frequency of high fire occurrence events was low, it may dramatically affect the analyses and modeling on the relationships between fire occurrence and a specific environmental factor. GWQR indicated that the relationships between forest fires and environmental factors significantly varied across the study area at different quantiles of fire occurrence. Compared to the global quantile model, the GWQR models performed better in model fitting and prediction at all quantile levels of forest fire risk. Therefore, the GWQR models provided a more complete view of forest fire distribution and highlighted the high risky locations of forest fires across the study area.

In addition to the forest fire count, the proportion / rate of forest fire occurrence was also evaluated and predicted in the third study. The response variable, the proportion of forest fire in an interval of 0 and 1 was calculated using the empirical data from year 2001 to 2016 across the study area, including both temporal and spatial information. To estimate how relevant topographical, meteorological, human, vegetable coverage, and land coverage factors influenced spatially and temporally on the proportion of fire occurrence, we applied the global beta model and developed the geographically weighted beta regression (GWBR) models, which were theoretically different from the traditional techniques of logistic regression and multiple linear regression. Our results showed that, generally, the likelihood of forest fire was higher in lower elevation, stronger sunshine (meaning smaller aspect index), denser settlement, and less cropland coverage. Other factors had different influence across the study region. Also, the environmental factors were spatially variedly related to the proportion of forest fire occurrence. In terms of model fitting and prediction, the global beta and GWBR models displayed better and more reasonable than the classical GWR model for the rate / proportion response variable. GWBR is an appropriate method to analyze the proportion of forest fire occurrence within a range between 0 and 1, and support better understanding for local prevention and management of forest fire.

Overall, this dissertation explored the spatially varying relationships between forest fire occurrence at both average and different quantile levels and related environmental factors in Fujian province, China, using both global and geographically weighted generalized linear models and quantile regression. As a global trend, the observed forest fire in this study was highly likely to occur in lower elevation, smaller aspect index (meaning stronger sunlight), heavier precipitation, smaller population density, less vegetation, wider grassland, and/or less cropland. Other environmental factors variedly relationships with the forest fire occurrence, including slope, relative humidity, temperature, road density, river density, settlement density, forest cover, and shrub cover density. The significant hotspots of predictor variables for forest fire count at both average level and different quantile levels, and the proportion of forest fire occurrence were evaluated respectively across the geographical locations of the study area. Therefore, the localized spatial models presented a more complete view of forest fire distribution and highlighted the risky local factors across the Fujian province, China. Hopefully, the information would assist the government agencies to make better decisions on where and what the fire management and prevention should be focused on with reduced economic expenses and improved the efficiency of forest fire management.

Curriculum Vita

Qianqian Cao

Ph.D. Candidate

Department of Forest and Nature Resources Management
College of Environmental Science and Forestry
State University of New York (SUNY-ESF), Syracuse, NY 13210, USA
Mobile: (315) 480-1106   Email: qcao03@syr.edu

Education

Ph.D. Forest and Nature Resources Management. College of Environmental Science and Forestry, State University of New York. Spring, May, 2020 (expected).
M.S. Applied Statistics. College of Arts & Science, Syracuse University. May, 2020 (expected).
M.S. Forest and Nature Resources Management. College of Environmental Science and Forestry, State University of New York. May, 2015.
B.S. Forestry. Beijing Forestry University, P.R. China. June, 2007.

Research Interests

Spatial statistics and modeling, general and generalized linear models, machine learning, data and text mining.

Computer Skills

Skilled and experienced programmer for SAS, R, SPSS, Matlab, Minitab, Solver; ArcGIS, SQL, Weka, Sci-kit Learn, NLTK, Mallet, Visio, Access.

Teaching Experience

Adjunct Instructor. August 2019 - Present. Department of Mathematics, Clarkson University.

- STAT 282 - General Statistics (3 cr, Fall 2019). Teach lectures, design Homework, Labs and Exams, and grade Exams.
- STAT 383 - Biostatistics (4 cr, Spring 2020). Teach lectures and computer labs.

Teaching Assistant. 2013 - 2018. College of Environmental Science and Forestry, State University of New York.

- APM 391 - Introduction to Probability and Statistics (3 cr). Taught lectures and computer labs, graded HWs and Exams.
- APM 510 - Statistical Analysis (3 cr). Taught lectures and computer labs, graded HWs, project reports, and Exams.

Journal Publications

Cao, Q., L. Zhang, Z., Su, Wang G., and F. Guo. 2020. Exploring spatially varying relationships between forest fire and environmental factors at different quantile levels. *International Journal of Wildland Fire*. Available at http://doi.org/10.1071/WF19010

Su, Z., M. Tigabu, Q. Cao, G. Wang, H. Hu, and F. Guo. 2019. Comparative analysis of spatial variation in forest fire drivers between boreal and subtropical ecosystems in China. *Forest Ecology and Management* 454. Available at http://doi.org/10.1016/j.foreco.2019.117669

Zhen, Z., Q. Cao, L. Shao, and L. Zhang. 2018. Global and geographically weighted quantile regression for modeling the incident rate of children's lead poisoning in Syracuse, NY. *International Journal of Environmental Research and Public Health* 15(10):2300. Available at http://doi:10.3390/ijerph15102300

Cao, Q., L. Zhang, Z., Su, Wang G., and F. Guo. Comparing four techniques to explore factors governing fire occurrence in Southeast China. Under review in *Ecological Indicators*.

Conference Presentations

Cao, Q. Modeling land-use proportions by beta regression and Bayesian beta regression. Presented at Northeastern Mensurationists Organization Annual Meeting. Stowe, VT. November 16, 2015.

Working Experience

Adjunct Instructor. Clarkson University, New York State, U.S. 2019-Present.
Teaching Assistant. College of Environmental Science and Forestry, State University of New York, New York State, U.S. 2013-2018.
Environmental Assessment Consulter. Sichuan Tianyu Environmental Assessment Company, Sichuan Province, P.R. China. 2007-2013.
Research Assistant. Beijing Forestry University, P.R. China. 2006-2007.
Forestry Integrated Practices of Silviculture. Fire Ecology and Management, Resources Management, Forest Survey and Planning, Heilongjiang Province. P.R. China. 2006.
Forestry Measuration Practice. Jiufeng Forestry Center. Beijing, P.R. China. 2006.
Ecology Observation. Xiaolongmen National Forest Park. Beijing, P.R. China. 2005.