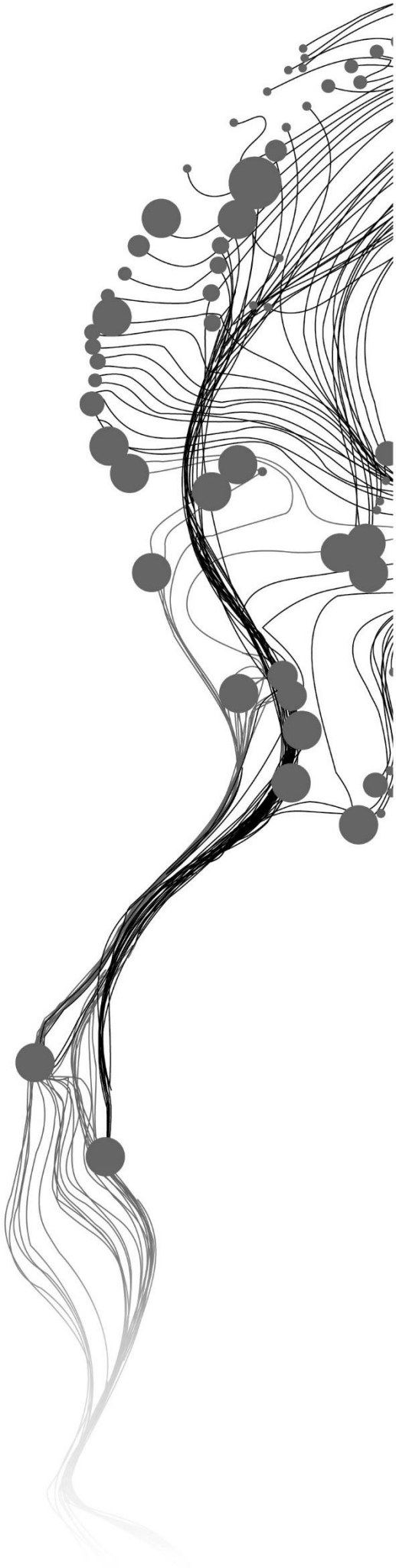
INVESTIGATING SPATIAL BIAS IN CITIZEN SCIENCE PHENOLOGICAL DATA

XINYI QIU
June, 2020

SUPERVISORS:
Dr. F.O.Ostermann
Dr. F.-B. Mocnik



INVESTIGATING SPATIAL BIAS IN CITIZEN SCIENCE PHENOLOGICAL DATA

XINYI QIU
Enschede, The Netherlands, June, 2020

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. F.O.Ostermann
Dr. F.-B. Mocnik

THESIS ASSESSMENT BOARD:
Prof.dr. M.J. Kraak (Chair)
Dr.ir. T.A. Groen (External Examiner, ITC Department of Natural Resources)

ABSTRACT

Volunteered phenological observations collected by citizen scientists are an important source of information for phenological studies. These observations are a type of volunteered geographic information (VGI). VGI has considerable value to various scientific research but contains inherent limitations. In phenology, the spatial bias in the distribution of observations may bring uncertainty to the representativeness of the observed geographic phenomenon. Spatial bias with respect to the uneven spatial distribution of observations is the focus of this study. The objective of this study is to identify and quantify the causes of spatial bias in sampling design and the influence on the spatial pattern of collected data. The variation in volunteer behavior and social-economic background which influence data collection is assessed through a statistical model. A point process model is adopted to model the relationship between observation intensity and a series of spatial covariates. The model primarily focuses on the first-order property of a point process – intensity, which only influenced by spatial covariates. The further analysis of residuals enables the visual interpretation of an unusual spatial pattern. The data is extracted from a national phenological network in the USA from 2003 to 2019. Seventeen geographic variables (human population density, road density, 15 different land cover classes) are included in the model. In general, road density and human population density are demonstrated significantly affect the collection of observations. Also, some particular land cover types show a higher probability for volunteers to make observations, such as the mixed forest of deciduous and evergreen trees, open space, and high-medium intensity developed areas. Overall, point process modelling shows a useful framework to analyse the spatial bias in VGI. The results demonstrate the effects of such bias, which can provide guidance for future volunteered data collection.

Keywords

Volunteered geographic information, citizen science, phenology, spatial bias, Poisson point process model

ACKNOWLEDGEMENTS

Foremost, I deeply thank my best parents for all the time unconditionally supporting my dream and encouraging me to go through all the tough days. I also wish to thank my grandparents, my sister for taking care of me all the years.

I wish to show my deepest gratitude to my first supervisor Dr. F.O.Ostermann for his valuable comments, unremitting guidance, and encouragement throughout my thesis. Without your continuous and illuminating guidance, this thesis would not be possible. My special thanks to my second supervisor Dr. F.-B. Mocnik for inspiring me and providing me useful and constructive suggestions. Thanks to my supervisors for making me stronger. I really learned a lot from you.

I would like to thank my friends Shan, Hao for helping me during the proposal. Thanks to these wonderful people make me a colourful life in Enschede.

Special thanks to my friends Xuechun, Xueting, Jiamei, Roubao in China. Without you I could sleep earlier at night.

Thanks to myself for never giving up.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background and motivation

Today, the increased popularity of GPS devices, information, and communication technologies have contributed to a large amount of Volunteered Geographic Information (VGI). VGI was originally termed by Goodchild (2007) and was defined as the geographic information that citizens voluntarily collect and share by means of digital tools. Nowadays, the use of VGI as source data for various types of scientific research is not a new notion. Many scientific projects using VGI as raw data have involved in multiple fields, ranging from social science (Sagl, Resch, & Blaschke, 2015), environmental monitoring(Joly, Vrochidis, Karatzas, & Karppinen, 2018), aquatic environment (Millar, Hazell, & Melles, 2019), disaster management (Ostermann & Spinsanti, 2011), biology (Zhu et al., 2015; Żmihorski, Sparks, & Tryjanowski, 2012), to phenology (Brunsdon & Comber, 2012).

Phenology is the discipline that researches " periodic events in the life cycles of animals or plants influenced by the environment such as weather and climate" (Cleland, Chuine, Menzel, Mooney, & Schwartz, 2007). Over the decades, numerous volunteered phenological observations have permitted scientists a novel way of accessing and creating detailed local knowledge to study phenological patterns. Taking advantage of volunteered phenological observations, scientists develop statistical models to monitor phenological responses to climate change. Also, volunteered observations can play a role in calibrating satellite data to better understand the relationship between field observations and remote sensing monitoring (Wallace et al., 2016).

The involvement of citizens in collecting observations have been proven to provide high-quality data for phenological studies (Beaubien & Hamann, 2011). With basic training, volunteers can provide reliable observations. Recent work on assessing voluntary phenology observations is mainly focused on volunteers' detection ability (McDonough MacKenzie, Murray, Primack, & Weihrauch, 2017) or consistency of recorded onset date (Mehdipoor, Zurita-Milla, Rosemartin, Gerst, & Weltzin, 2015). Apart from such quality issues that affect the accuracy of the collected data, bias in the spatial distribution of collected observations is less explored. Spatial bias refers to the uneven spatial coverage of observations in a defined area. The existence of spatial bias in volunteered phenological observations has been acknowledged in several studies. Miller-Rushing et al. (2008) found out that the first reported flowering date is strongly correlated with the sample size of observations. They controlled the number of observations in three types – increasing, unchanged, declining, and found out that the more observations (the bigger the sample), the earlier (more extreme) the first reported flowering date. A similar situation exists in the research by Brunsdon and Comber (2012). They discovered more extreme values in regions in which more samples are provided, which made it difficult to determine whether the occurrence of extreme events is exceptional or common. Phenological models largely rely on the accuracy of recorded data, especially the data with extreme values, as extreme early or late observation in phenology implies the variation under the influence of climate change. However, the number of observed extreme events can be influenced by the spatial distribution of observations. Therefore, the real spatial distribution of extreme onset dates may be uncertain, thereby affecting the accuracy of generated predictive models.

There are certain studies focus on mitigating the influence of spatial bias by removing samples based on specific criteria (Boria, Olson, Goodman, & Anderson, 2014; Varela, Anderson, García-Valdés, & Fernández-González, 2014). However, such approaches improve the predictive models but may result in the loss of key information. Considering the heterogeneous nature of VGI, the data collection process can be opportunistic if it lacks standardized protocols. This poses challenges for achieving an expected spatial distribution trend. In general, the number of observations varies in space can be attributed to spatial heterogeneity – variation in local environment factors and social-economic characteristics, etc. Existing analysis found volunteers are more likely to make observations near where they live or at more attractive places (Dennis & Thomas, 2000). Or more observations are reported near roads or in areas where public transportation is more convenient (Mair & Ruete, 2016). These studies investigated the potential causes of spatial bias of observations and can be further explored to provide guidance for the future observation process.

Current guidance for phenological observations is generalized and is in the form of simple descriptive suggestions (Koch et al., 2009). And the guidance for making phenological observations gives large freedom to volunteers to choose the observation site. Taking spatial data related to volunteer effort into account is helpful to understand the interaction between observation patterns and the data collection process. Therefore, through identifying spatial bias during the data collection process, we can provide better guidance in effective volunteer management and improve decision-making in future data collection design.

1.2. Research identification

Spatial bias concerning the distribution of observations has become an issue of concern in assessing the data quality of VGI. Volunteered phenological observation as a special type of VGI also has such quality issue. Although the effects of spatial bias in volunteered phenological observations have been acknowledged, few have attempted to evaluate the causes of spatial bias. Therefore, this study investigates the potential causes of spatial bias from the perspective of the data collection process, which can in turn help guide volunteers to collect observations. On the basis of Geldmann et al. (2016), this study conducted a more detailed analysis of the relationship between observation intensity and potential spatial factors.

1.2.1. Research Objectives

The overall objective of the study is to understand and quantify how bias in data collection procedure influences the spatial pattern of observational data, and to enable guidance for future data collection. To achieve this goal, the sub-objectives are:

1. To identify the spatial bias in the volunteered observational data.
2. To identify the factors that determine the spatial bias in the data collection procedure.
3. To adopt a method to estimate the causes of spatial bias on the spatial pattern of observations.

1.2.2. Research Questions

This research attempts to address the following research questions with regard to each sub-objective.

1.1 What is the spatial pattern of volunteered phenological observations?

1.2 What is the spatial bias in volunteered phenological observations?

2.1 Which spatial factors lead to the spatial bias of volunteered observations?

3.1 How to identify the influence of these factors on the variation of the spatial distribution of volunteered observations?

3.2 To what extent these factors explain spatial bias?

1.3. Thesis structure

The thesis is composed of six chapters. Chapter 1 describes the research background, problem identification, objectives, and associated questions. In Chapter 2, the literature review gives a brief explanation of current works on the exploration of spatial bias in VGI, as well as related work to the research topic and adopted method. Chapter 3 introduces the data materials used for the research, and methods adopted to answer relevant research questions. Chapter 4 presents the results derived from the adopted methods. Chapter 5 discusses the explanation of results, as well as limitations and recommendations. Chapter 6 summarises the entire work of this study and advises on future research.

2. LITERATURE REVIEW

The review of related work on spatial bias in volunteered geographic information (VGI) justifies the importance of conducting such analysis that can be applied to give guidance in the collection of VGI and then help improve the quality of VGI. According to this, this chapter firstly gives a general introduction to VGI. Second, the context of VGI related to spatial data quality help to understand the use of VGI and associated issues. Third, the spatial bias in VGI quality issues provides previous research close to this study. At last, the applications of the method used in this study are introduced.

2.1. Volunteered geographic information

This section introduces the origin of VGI and the usability in VGI as a novel data source in scientific research as well as its characteristics.

Working with citizen scientists dates back to 1900, the National Audubon Society's annual bird count (Cohn, 2008), where volunteers assisted in data collection to provide information about the trends of bird species. In recent decades, the evolving advanced communication technology and GPS-based devices have enabled volunteers to act as sensors to provide valuable information about what they sense at a local scale, assisting scientists in data collection. Furthermore, the development of smart applications has facilitated many geographic citizen science projects to produce geo-information products (USA-NPN National Coordinating Office, 2016).

VGI is known as a phenomenon that the public voluntarily contributes to geographic information in a direct or indirect way (Goodchild, 2007). The involvement of volunteers in monitoring geographic phenomenon is of great value to understand the earth's surface. Volunteers can provide more efficient local knowledge than experts for creating valuable information. OpenStreetMap is such a best-known project where many users engage in mapping and contribute to geospatial information they sense at the local scale. In addition, VGI can provide information with more details. For instance, in phenology, satellite images offer valuable information in monitoring the seasonal cycle of plant species, but usually at large scale, while VGI-based ground observations can focus on the changes of particular life stages of individuals (Melaas, Friedl, & Richardson, 2016). Satellite images usually contain multiple individuals and diverse land cover in a large pixel, but the integration with VGI can help understand the phenological status of the individual organism. Furthermore, VGI with high quality can be used as a calibration tool to complement traditional satellite images, thereby improving the accuracy of land cover products (See et al., 2016).

Citizen science is defined as an activity that non-professional citizens actively and voluntarily help collect observational data for scientific projects(Cohn, 2008). Geographic citizen science projects which include the collection of geographical information contribute to VGI (Haklay, 2013). In recent years, many geographic citizen science projects related to environment and biology have been developed (Chandler et al., 2017; Fritz, Fonte, & See, 2017). These types of citizen science projects provide key information on monitoring the earth's surface, and the subsequent processing and analysis of VGI facilitate to understand the geographic phenomena. For instance, volunteers have assisted in monitoring the growth trends of animal and plant for several decades ("USA National Phenology Network," n.d.; "Nature Today | De Natuurkalender," n.d.). The data is then used to build observational networks for studying the spatial variation of natural phenomena. Subsequently, further products based on observations are used to study the impacts of climate changes on animals and plants. Similarly, species monitoring projects (e.g., eBird)

(Sullivan et al., 2009) observe species occurrence to better understand the ecosystem's health and effects of climate change. With the help of limited representative samples collected by volunteers, some geographic phenomena (e.g., habitat suitability) that are not easily observed can be produced by other techniques (Zhu et al., 2015).

However, VGI is not always ideal. There has been a growing concern about using VGI as a data source for various scientific projects. An important concern of VGI is the spatial data quality when using VGI as inputs of scientific researches. It is worth noting that the efficient use of VGI depends on the representativeness of collected samples and the quality assurance of data (Zhang & Zhu, 2018). The following section will introduce quality issues in VGI.

2.2. Spatial data quality of Volunteered geographic information

Due to the considerable potential use of VGI, the quality of VGI is always a subject of much concern, in order to acquire valuable and accurate information for decision-making. However, the heterogeneous nature of VGI, arising from the lack of specification and standardization in the process of creation, has introduced uncertainties into the quality of VGI.

The data quality is an important consideration, it directly influences the efficient use of VGI. The data quality of VGI is often described by five aspects: completeness, positional accuracy, temporal accuracy, thematic accuracy, and consistency (Antoniou & Skopeliti, 2015; Senaratne, Mobasheri, Ali, Capineri, & Haklay, 2017). Accordingly, many works based on these five aspects have been conducted to assess and assure the data quality of VGI. Completeness related to the absence of features in a dataset has been explored in many works (Jackson et al., 2013; Camboim, Meza Bravo, & Sluter, 2015; Jacobs & Zipf, 2017). Commonly, a comparison method is used to assess the feature completeness. Their findings showed the incompleteness of a dataset result from the uneven coverage of volunteer population and imbalanced volunteer effort. As for other aspects, Mehdipoor et al. (2015) checked the inconsistent observations by integrating observational data with several environmental contextual information. Camboim et al. (2015) compared features in OpenStreetMap with several indicators to identify changes in temporal scale and to check the temporal quality. The positional accuracy of volunteer-generated maps is usually evaluated by comparing with official data such as satellite images (Al-Bakri & Fairbairn, 2010). Similarly, thematic accuracy with respect to attribute accuracy of OpenStreetMap was assessed by comparing with high-resolution reference data (Arsanjani et al., 2015). Although the assessment of VGI data quality with respect to theses five aspects has long been explored, these quality indicators are not absolute. The utility of different types of VGI differs for specific purposes, so it needs the evaluation of VGI data quality on a case-by-case basis (Feick & Roche, 2013).

In addition to evaluating VGI quality at the level of product, many studies pay attention to the potential bias in the data collection process. The collection process of VGI can be "opportunistic" because the collection of target samples lacks specific surveying design and standard protocols (Van Strien, Van Swaay, & Termaat, 2013). Accordingly, several studies focus on regulating specifications during the production process were proposed in order to address quality issues. Brando and Bucher (2010) proposed quality metadata that is of specifications for users when producing spatial data, which can help improve the data quality. Moreover, different levels of contributors with varying behavior, awareness, and preference may result in spatial bias of collected observations. For example, volunteers may have a tendency to report observations that they show more interest in. Vyron and Schlieder (2014) pointed out that volunteers tend to choose the areas and features which they show preference to, which leads to the bias in the spatial distribution of contribution and affects the completeness of the dataset. Besides, variation in the intensity of data collection at different time periods would influence some specific projects. Volunteers are more

likely to collect more samples during holidays than on weekdays. Żmihorski, Sparks, and Tryjanowski (2012) found that there is a weekend bias in the occurrence of observations. Volunteers tend to report more observations during weekends than weekdays, which will bias the estimated results especially when the research interest is the relationship between species distribution and climate conditions, because some climate conditions are different on weekends.

Regarding biases in the data collection process, Van Strien et al. (2013) summarised four types of spatial bias which are raised by the variation in observer activity: (a) geographical bias (unbalanced spatial coverage in a given area) (2) reporting bias (varying tendencies for volunteers to report certain species) (3) observation bias (varying sampling effort) (4) detection bias (uneven ability to detect all target phenomena). These biases will lead to the unrepresentativeness of valuable observations, thereby resulting in biased estimates. Especially, in phenology study, Van Strien et al. (2008) pointed out that the number of reported earlier onset date of plants increased with the increasing number of observations at a place, which may result in biased estimation of phenological changes. Under ideal circumstances, the initial experimental design expects a uniform spatial coverage of sampling over the entire study area. However, the data collection process is "opportunistic", the real distribution of collected observations is spatially heterogeneous. For one thing, the spatial heterogeneity of the distribution of observations may be due to the biased geographic background. For example, population differences among urban and rural areas would affect the coverage of sample collection, as densely populated areas may increase the abundance of observations. Thus, such population characteristics are often used in exploring biased patterns of spatial distribution (Li, Goodchild, & Xu, 2013; Camboim, Meza Bravo, & Sluter, 2015; Zhang & Zhu, 2018). Population density is served as additional data to assess the intensity of observation, which can help check the richness and completeness of the observed entity at a spatial location. Camboim et al. (2015) investigated the correlation between the updated data collected by volunteers and the urban-rural population to assess the completeness of collected features. Their findings showed the number of collected samples is highly correlated with areas of high human population density. For another, the spatial heterogeneity is reflected by the observer heterogeneity (Mocnik et al., 2018). The volunteer behavior of data collection may be affected by the surrounding environment. For example, volunteers may be only interested in collecting data close to where they live, which may result in the unrepresentative location of the observed sample. Kadmon, Farber, and Danin (2004) found that road density and accessibility affect the intensity of observation. The results showed the frequency of observations close to roads was much higher than that expected from a uniform coverage. Mair and Ruete (2016) also found that observations are more likely to occur in areas that are easier to access by public transportation.

Other forms of spatial bias are also discussed in some studies. In an eBird project, volunteers are provided with customized routes to make observations, while they may plot location point without standardized format. As a result, observations are intensively concentrated at the beginning and end of the route rather than at the exact location where they occur (Sullivan et al., 2009). In such cases, unrepresentative and biased samples adversely affect the results deduced from VGI. Also, spatial patterns of a geographic phenomenon predicted from biased data may lead to wrongly recognize the patterns, ignoring minority patterns with meaningful information (Basiri & Gardner, 2017). Basiri and Gardner (2017) explored the impact of biased data in an actual case. They intentionally added a different degree of bias into original data in order to find out how bias data influence the data mining process and model results. The result shows that less-biased data can provide more reliable information than very-biased data.

The spatial bias in the observational data would largely become barriers to the use of VGI and may restrict the usefulness of inferences drawn from the data. It is therefore needed to investigate the cause of the spatial bias so as to optimize the data collection process to collect data more effectively.

2.3. Spatial bias in volunteered phenological observations

This section introduces the potential spatial bias in volunteered phenological observations and its potential causes in the data collection process. In this study, spatial bias refers to the uneven spatial coverage of the observations in the entire study area.

The spatial coverage of observations at a defined area is one particular focus of data quality of volunteered observations, which is related to the completeness issue. Yesson et al. (2007) compared the occurrence data of species from the volunteered collection with the official taxonomic database. They expected uniform distribution of 5400 species across the whole region. But they found around 700 species lack valid observations because of the problematic location information, which results in the uneven distribution of species in space. Although more than 80% of the survey sites had been proved accurate, the uneven distribution of species may lead to misleading of distribution of species richness. Morellato et al. (2009) found that the lack of sufficient observations would lose information for phenological events. Also, if a large number of observations were recorded for only one species, the results would overestimate the presence of that species.

In general, data collection relies on volunteers. Some authors used the estimation of sampling effort by volunteers to identify spatial bias. Mair and Ruete (2016) examined the number of observations in a grid cell and compared the effects of several geographical variables that explain the spatial bias in the distribution of observations. They found the influence level of each geographic variable was varying according to different species. Road density proved to have the highest correlation with variation in observations. However, the effect of road density did not increase monotonically with the increase of the number of observations, they argued the observer's access to a site through different vehicles is also a potential factor. Another similar result was presented by Reddy and Dávalos (2003). They demonstrated that sampling was significantly biased along the road in the data collected. They also tested the distribution of observations inside and outside the nature reserve areas. The result showed sampling was preferred near nature reserve areas, which could result in sampling artefacts of the true species distribution.

2.4. Point process model

Point process models (PPMs) have been widely used in species distribution modelling (SDM), where SDM is used to predict if the spatial distribution of a species is dependent on associated environmental and geographical factors. PPMs regard the distribution of species as a set of point events, where the point events represent the presence of certain species. In general, there are two interests in PPMs, the prediction based on the observed pattern and the explanation of the relationship between observed pattern and associated environmental conditions (Renner et al., 2015).

Furthermore, the selection of a proper model depends on the interested property of a point process. The fundamental properties of a point process are (i) intensity —the expected number of events per unit area (ii) inter-point interaction. This study focuses more on the intensity function. A common model that is entirely determined by the intensity function is the inhomogeneous Poisson process model. In the model, the intensity is spatially varying, the distribution of points is determined by the intensity function. To understand the inhomogeneous Poisson process, the intensity of points can be defined by introducing the relation of covariates to the point events. Covariates are data used to explain the intensity and can be considered as explanatory variables. Covariates need to be spatially continuous across the entire area, so covariate information is defined at every point in the area (Illian, Penttinen, Stoyan, & Stoyan, 2008).

Reinhart and Greenhouse (2018) proposed a spatial-temporal PPM that incorporates spatial covariates to predict the distribution of crime. They assumed the occurrence of a crime may cause a repeat offender in the near future. Thus, a former crime event at a place can be served as a spatial covariate of a future crime event. They also used several spatial covariates that potentially promote crime to predict a future crime event. Their results showed that this model can predict future crime events from both spatial and temporal aspects.

Hefley et al. (2013) applied PPMs to estimate species distribution and investigate detection bias during sampling. A marked PPMs was fitted to test the species richness where the sample size of a species was treated as the mark. A mark in PPM is the additional information that only associated with a point, and is considered to affect the intensity of points (Illian et al., 2008). They argued that additional information like environmental features is necessary for studying the effects of spatial bias on the predictive distribution of species.

Since PPMs predict species distribution with the function of spatially varying covariates, it also provides insights into explaining the effect of each covariate. Niemi and Fernández (2010) proposed a Bayesian PPM approach to simulate the variation in animal density. The used a single covariate as the function of the animal abundance. They argued the uncertainty of density variation is not only the effects of covariate but the effects of areas without sampling. Similarly, Geldmann et al. (2016) applied a PPM to estimate the effect size of each covariate on the intensity of observations. Moreover, both these two studies used residuals to indicate the observation intensity under the influence of spatial covariates. The residuals analysis helps understand the changes in the distribution of target objects cause by covariate settings.

3. DATA AND METHODS

This chapter provides detailed information about the study area, and data applied in the subsequent analysis, and also includes data processing methods and adopted data analysing methods that aim to answer research questions.

3.1. Study area

Close to 40% land of the eastern United States is covered by forest (Delcourt & Delcourt, 1996). The forest is dominated by deciduous trees. These forests are "temperate forest" which have a visible response to the changes in temperatures and precipitation (U.S. National Park Service, 2017). Scientists have used seasonal leaf development to develop phenological models to explore the plant reflection of climate change for years (Melaas et al., 2016). The selection of the study region considers the main temperate forest cover. The United States Environmental Protection Agency (EPA) defined ecological regions of the entire North America, where the eastern temperate forest ecological region of the United States is defined between around 67°W-95°W and 24°N-48°N (United States Environmental Protection Agency, n.d.) (Figure 1). The common tree species of the eastern temperate forest include oaks, maples, hickories, and pines. The eastern temperate forest ecological region is distinguished by suitable temperature and moisture and has four distinct seasons, which provides an ideal living environment for the deciduous forest.

The study region encompasses nearly 33 states. The most densely populated cities are New York, Philadelphia, Washington. Taking advantage of the support from the government, numerous citizen science projects are in full swing. The U.S. government encourages the public to join in scientific research by providing an official portal where more than 400 citizen science projects are active (CitizenScience.gov, n.d.). Moreover, the urbanization process has promoted diversification in the eastern region. Considering the need for a sufficient user base and representative regional landscape to explore spatial variation in sampling effort, we choose the eastern United States as the study region.

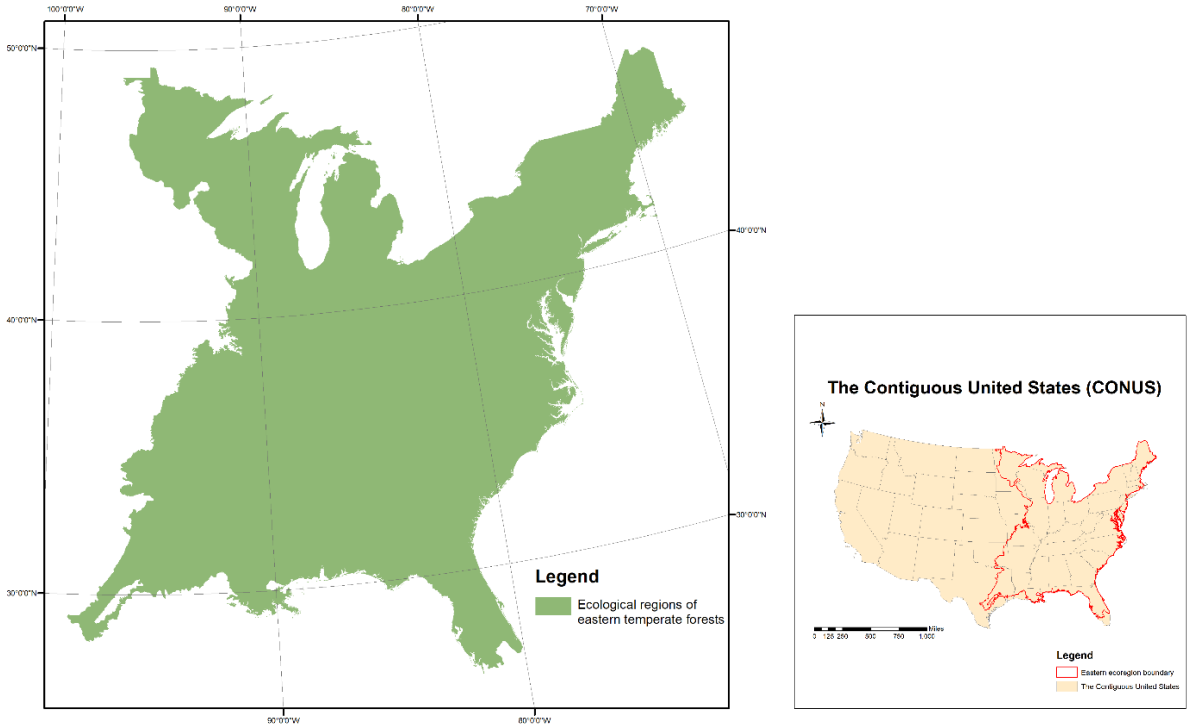


Figure 1 The overview of the study area

3.1.1. Phenological VGI datasets

Our specific problem of interest involves the exploration of spatial bias of volunteered geographic information. In this research, spatial bias is the reflection of sampling bias, which focuses on volunteers' unbalanced effort to detect a target object and thus lead to uneven coverage of samples in space. In order to evaluate spatial bias in volunteered geographic information, we used a highly referenced phenological observation dataset which has proven high reliability and utility (Fuccillo et al., 2015; Rosemartin et al., 2015; Mehdipoor et al., 2015).

Volunteered phenological observation data were derived from the USA National Phenology Network (USA-NPN), a national phenology program that records long-time volunteer observations of phenological events (USA-NPN National Coordinating Office, 2016). More than 11 million observations of plant status and over 1200 species have been recorded across the continental U.S. since 1981. The program provides volunteers with training in skills and particular equipment to report observations. As part of their mission, data quality assurance has been implemented to avoid misidentification in species and phenophase status. Observers record the phenophase status they observe each time by reporting yes/no/uncertain of a defined phenophase. As described in the official guidance on site selection (USA National Phenology Network, n.d.-a), site selection is largely dependent on observers' preference. There are no specific rules for site selection, but several general guidelines. Any independent volunteer can add a preferred location as a fixed observation site, such as a backyard or natural park, as long as the site is as representative of the surrounding environment as possible, and is similar to surrounding forest habitat. The size of a site for observing plants does not matter much, it depends on the convenience of observers. If several individual objects (e.g. several trees of the same species or several species) are observed, it can be considered at one site when individuals grow under similar site conditions and the site area is no larger than 6 hectares (250 * 250 m). It is suggested to select sites are easily accessible so that observers can visit often.

Particularly, there are two types of observers in the dataset, which are partner group observers and individual observers. A partner group is one where many observers contribute to observations at locations defined by the group administrator. Members have no permission to add sites. Partner groups in this dataset can be educational organizations, botanical gardens, university groups, etc. Based on the exploratory analysis, observations are spatially dispersal to some extent even if they belong to the same partner group. So in this study, we assume that observations from both partner groups and individuals are independent of each other.

These observational data from USA-NPN have been widely applied in various themes, especially in model construction for climate drives of phenology and prediction of phenological events. The particular species used in the research is based on predictive modelling of the growth stage of deciduous forests by Melaas et al. (2016). We extracted flowering records for 9 species that belong to deciduous broadleaf and include in the study region: (1) red maple, (2) sugar maple, (3) paper birch, (4)American beech, (5) quaking aspen, (6) black cherry, (7) black walnut, (8) white oak, (9) north red oak. Each observation was provided with an individual phenometrics dataset which encompassed estimates of phenophase onset dates, along with several ancillary datasets for information on observers, sites, and individual plants.

Data, which are represented as a set of points in space and have records of locations, are defined as point events (Diggle, 2013). A volunteered phenological observation with reported location can be regarded as a point event. A collection of point events that randomly located in an area is a spatial point process (Renner et al., 2015). In most cases, volunteered phenological observations are often recorded continuously for a long time. Thus one individual object may have multiple records at different times. But the site of an observed object is spatially static and unchanged. In this study, only the spatial dimension is considered, which means a long-term volunteered phenological observation is an individual point event. Therefore, an individual object recorded multiple times is screened one record.

The observation dataset contains 15825 observation records for 9 species reported at 1259 sites. One observation with multiple records was filtered as one individual observation and corresponding additional information is attached. The code shown below elaborates the data processing (Figure 2).

```
obs = pd.read_csv('individual_phenometrics_data.csv')
site = pd.read_csv('ancillary_site_data.csv')

# Screen out the required column names
obscols = ['ObservedBy_Person_ID', 'Partner_Group', 'Site_ID','Latitude',
           'Longitude','Individual_ID','Multiple_Observers']
sitecols = ['Site_ID', 'Site_Type']
obs=obs[obscols]
site=site[sitecols]

# Merge the table
obssite = pd.merge(obs, site, on='Site_ID',how='inner')

# Filter one observation per Individual_ID
ois = obssite.groupby(['Individual_ID','Site_ID']).head(1)
ois.to_csv('one_individual_site.csv',index=False)
```

Figure 2 Code for data filtering

In the code, table "obs" contains information about observers and record date and spatial location of an induvial observation. Table "obs" was merged with the table "site" based on the same "site ID". Each individual plant was selected for the first record. We finally screened out 2916 individual observations. Each individual observation contains the specific location reported by one or multiple observers. The site location of each observation is provided with accurate coordinates that have been geocoded by USA-NPN based on users' descriptions of address. Table 1 shows the description of the filtered dataset. Figure 3 displays the relation and the structure of the observation data.

Table 1 The overview of attributes

Attribute name	**Description**
ObserverBy_Person_ID	The unique identifier of each observer
Partner_Group	The name of the partner group which indicates an observation is monitored by a partner group; "-9999" indicates monitored by an individual observation
Site_ID	The unique identifier of the site
Site_Type	The type of site information on whether the site is managed by a group or an individual observer
Individual_ID	The unique identifier of an individual plant
Latitude	Site latitude
Longitude	Site longitude
Multiple_Observers	Indicate whether an individual observation is monitored by one or multiple observers

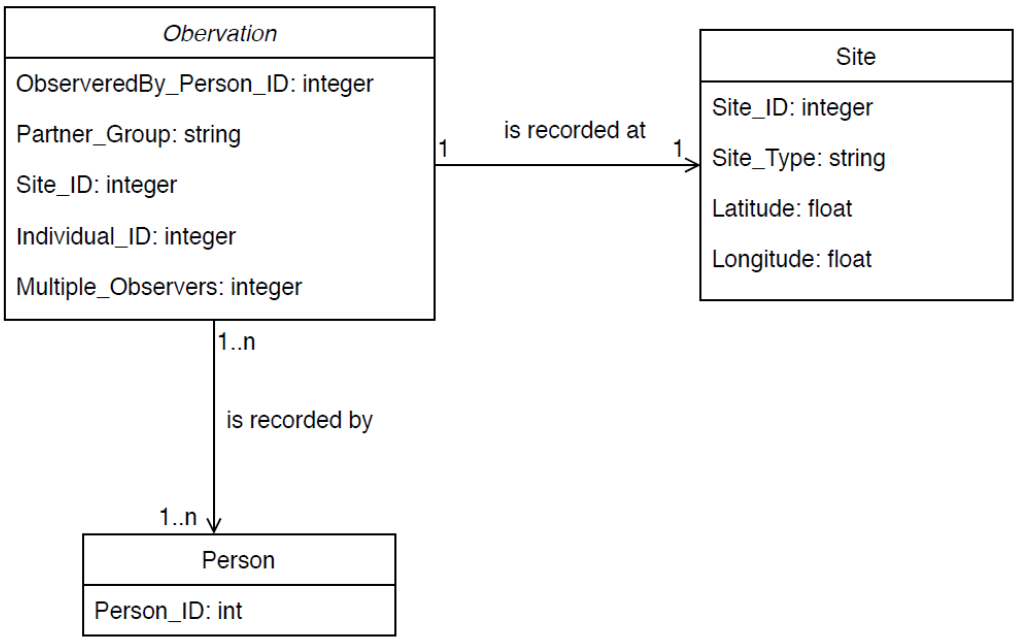


Figure 3 The conceptual data model that defines the compositions of an observation. Each observation with a unique identifier was recorded at a site. Each observation can be recorded by one or several observers, also, one observer/multiple observers can record one or more observations.

3.1.2. Data accounting for spatial bias

This section introduces the data used as spatial covariates to account for the function of observation intensity. The preprocessing of these data is elaborated in detail along with the description of data collection.

3.1.2.1. Population density

The U.S. population data was derived from NASA Socioeconomic Data and Applications Center (SEDAC). The raster dataset consists of demographic and socioeconomic data from national censuses. The initial population data and administrative units are derived from the database of U.S. Census Bureau which is a national agency that provides high quality and openly census data on socioeconomic (U.S. Census Bureau, 2014). The dataset contains the total population counts based on the irregularly shaped census block. The gridded population product is generated through assigning the counts to a regular quadrilateral grid, where each grid cell is allocated in proportion to each census block data when a grid cell has a mixture of two census blocks (SEDAC, n.d.). The spatial resolution is approximate 1 square kilometer.

The gridded population data was extracted based on state geography from the database. In order to merge all state data to fit the study area, a "Mosaic" tool in ArcMap was used. The 33 state-based raster datasets adjacent to each other were merged into one entity. The output of the image mosaic was then clipped to fit the study area. Figure 4 shows the data processing of human population density.

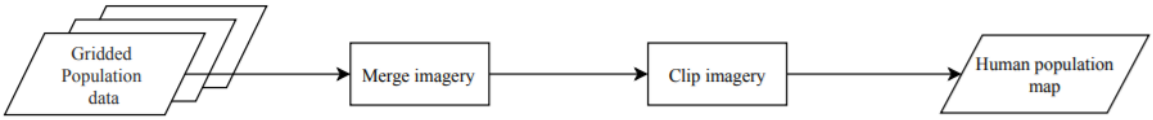


Figure 4 Flowchart for data processing of human population density

3.1.2.2. Land cover

In order to detect the impact of the different surrounding landscapes on spatial patterns of observations, the land cover dataset was used to study spatial bias. The land cover dataset was derived from the National Land Cover Database (NLCD), which is a national database that belongs to federal agencies and provides national land cover products with high quality. Integrated with geospatial ancillary datasets, the generation of the land cover product is based on a decision-tree classification algorithm using Landsat imagery (Yang et al., 2018). Some types of land cover not included in the study area were eliminated, and finally, 15 classes were used in this research, as shown in Table 2.

The spatial covariate is required to be spatially continuous across the study area for subsequent model construction, thus the land cover dataset was split into individual class and then calculated the coverage in each grid cell. The derived land cover dataset has a raster resolution of 30 m for the entire conterminous United States. We defined the coverage by calculating the areas of target land cover class within each grid cell. Cells with null values were replaced with zero value. Each grid cell has an area of 1 square kilometer. This could avoid the loss of information comparing with directly reclassing the data to the spatial resolution of 1 km. The range of the coverage is from 0 to 1. Figure 5 shows the data processing of land cover.

Table 2 The overview of the land cover class

Land cover class	Land cover value	Description
Water	11	Open Water
Developed	21	Developed, Open space
	22	Developed, Low intensity
	23	Developed, Medium Intensity
	24	Developed, High Intensity
Barren	31	Barren Land(Rock/Sand/Clay)
Forest	41	Deciduous Forest
	42	Evergreen Forest
	43	Mixed Forest
Shrubland	52	Shrub/Scrub
Herbaceous	71	Grassland/Herbaceous
Cultivated	81	Pasture/Hay
	82	Cultivated Crops
Wetlands	90	Woody Wetlands
	95	Herbaceous Wetlands

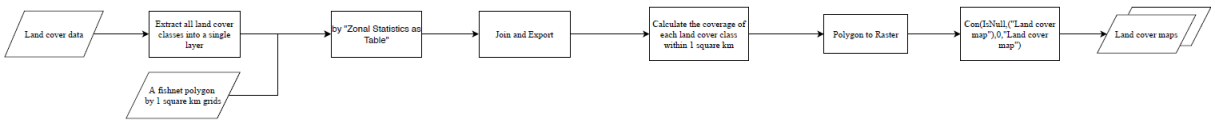


Figure 5 Flowchart for data processing of land cover

3.1.2.3. Road density

Sampling along the roads may lead to the complete omission in data collection in roadless areas. We assume roadside surveys may contribute to biased spatial patterns. The source dataset of road types was from the U.S. Census Bureau. The road density raster map was created by extracting several road types including primary roads, highways, bike paths, walkways/pedestrian trails, and calculating the total length of all roads in a 1 km grid.

The road density was calculated as the length of the road per grid cell and was processed in ArcMap. The road types were filtered from the U.S. Census Bureau's 2014 TIGER database to get all line segments and merged. A fishnet polygon with 1 square kilometer was created and intersected with all the line segments to get lines in each grid cell, which can be achieved by the tool "Identity". Then the attribute of line data was joined with the fishnet polygon based on the same unique ID. The fishnet polygon was converted to a raster map, but there are some pixels with null values. Thus the raster calculator was finally used to replace the null value with 0. Figure 6 shows the data processing of road density.

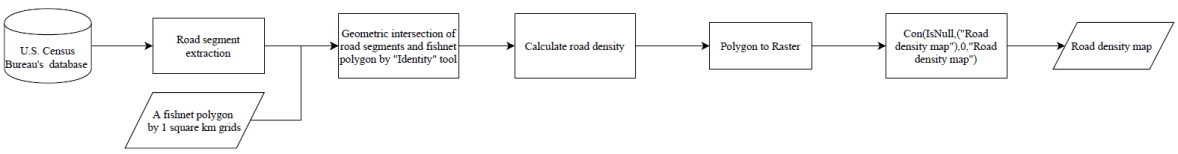


Figure 6 Flowchart for data processing of road density

3.2. Methods – modelling

The construction of a Point process model is the main focus of this study. First, this section describes data preparation for model construction. Second, the methods designed to answer research questions are presented. In the end, validation methods were given to confirm the model. The overall process is summarized in Figure 7.

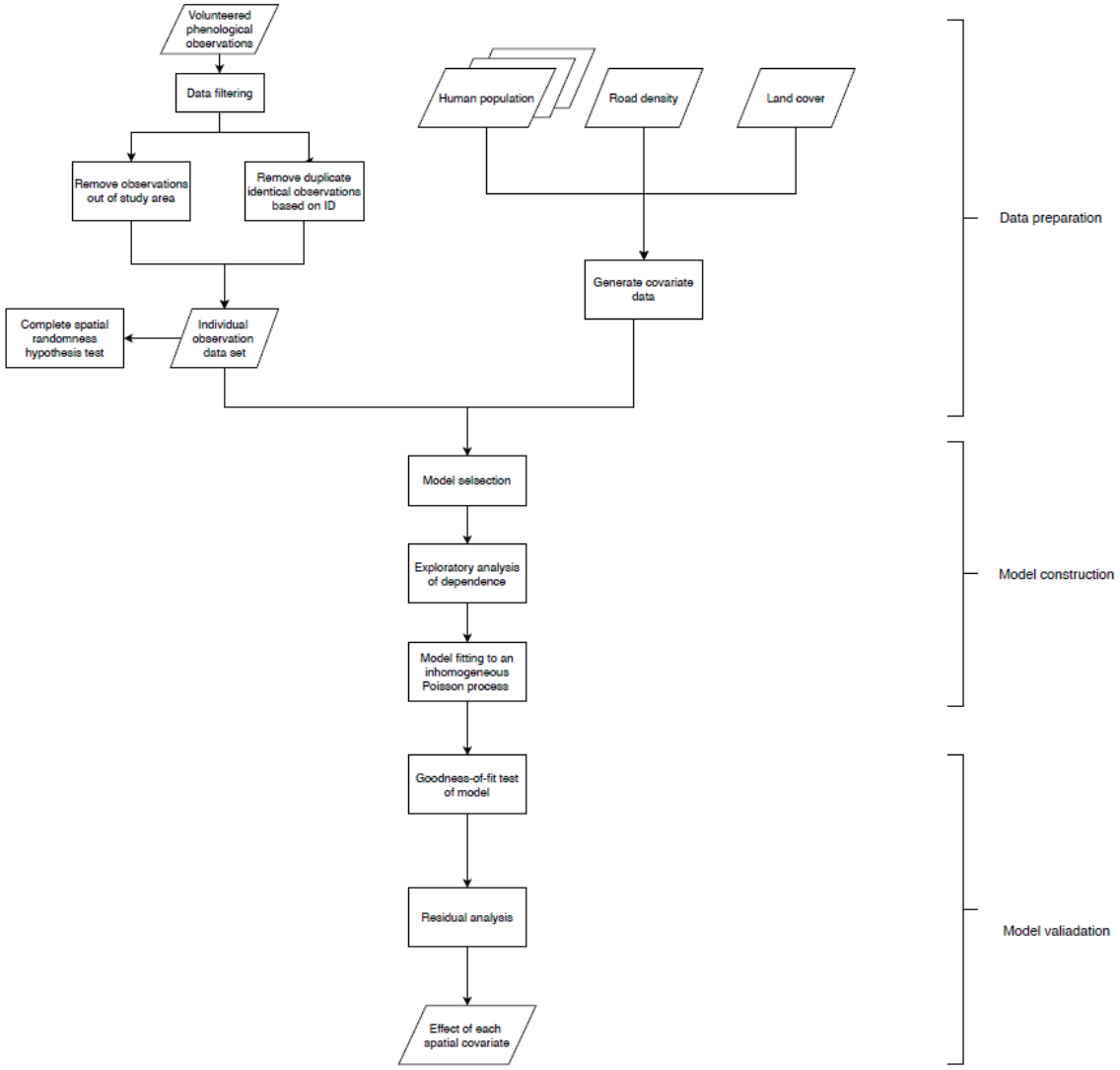


Figure 7 Overall implementation workflow

3.2.1. Complete spatial randomness

The first step before constructing a suitable model is an exploratory analysis of the specific data set. This step is to answer the research question 1.1.

A spatial point pattern is a set of points irregularly distributed in a one-, two-, three-dimensional plane (Diggle, 2013). Each point represents the location of an observation event. Figure 8 shows a point pattern that, provides the spatial location of the observation event occurring in a defined study region. The study region is defined as the "window" in mapping a spatial point pattern. The location of a point event is represented by Cartesian coordinates in the window. Volunteered phenological observations including

locations can be regarded as a stochastic point process. A point process is a stochastic model of irregular point patterns.

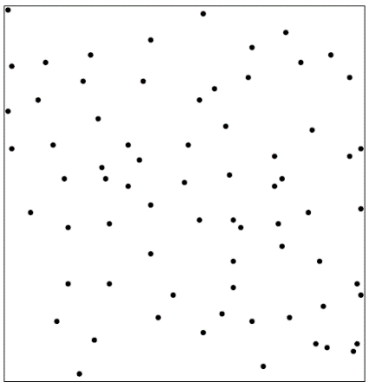


Figure 8 A two-dimensional point pattern (Baddeley, 2010)

Commonly, Complete Spatial Randomness (CSR) hypothesis test is implemented to identify the spatial point pattern. Furthermore, CSR hypothesis test can serve as a premise for the subsequent selection of a suitable class of models (Illian et al., 2008). CSR can be characterised as a homogeneous Poisson process. The properties of CSR are (Diggle, 2013):

(i) The number of points in any region A follows a Poisson distribution with mean $\lambda * \text{area}(A)$, where λ is the intensity of points per unit area.

(ii) the locations of point events are independent of each other.

Property (i) explains that the probability of each event occurring throughout the study area is the same, and the intensity λ is a constant and does not change over the study area. Property (ii) explains the complete randomness, which means that there is no interaction between points.

In the CSR hypothesis test, the "null hypothesis" is the homogeneous Poisson process, which means points are independently and uniformly distributed over the area. The null hypothesis is rejected when a large difference exists between the estimated summary characteristic and theoretical summary characteristics. The rejection of CSR indicates the potential relationship among point events, and the cause of departure from CSR provides guidance for other Poisson process models.

Quadrat counts are often used to construct hypothesis testing. In the Quadrat counts test, the entire study region is divided into equal subregions, or "quadrats". The number of point events that belong to the corresponding "quadrats" is then counted. Besides, the number of "quadrats" determines the outcome of the test. Further, in this study, the χ^2 test will be used for the goodness-of-fit of the uniformity hypothesis. The Pearson χ^2 test based on quadrat counts assumes the counts of each quadrat subregion are independent. The value of χ^2 relies on the deviation between the observed value and theoretical value. The Pearson χ^2 test statistic is defined by (Diggle, 2013):

$$\chi^2 = \sum_{i-1}^{m} \frac{(n_i - \bar{n})^2}{\bar{n}}, \tag{3-1}$$

when Pearson $\chi 2$ test follows the null hypothesis of the uniform distribution of point events, n_i denotes the observed intensity in each quadrat, $\bar{n}$ denotes the theoretical intensity in each quadrat, m denotes the number of quadrat regions. The intensity in CSR is calculated through diving the frequency of events in each subregion by the area of each region. The p-value is used to evaluate the test result. Figure 9 shows an example of the quadrat counts of a point pattern.

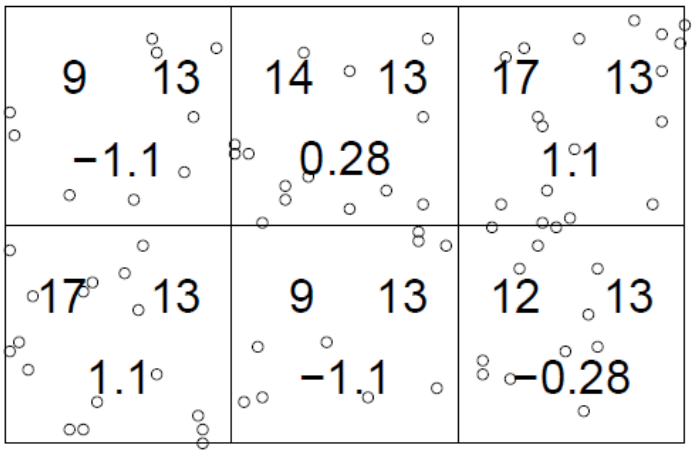


Figure 9 Quadrat counts of a point pattern. In each quadrat, the upper left indicates the number of observed point events, upper right indicates the number of expected point events, the bottom is the Pearson residual.

However, the result of Quadrat counts heavily depends on the selected number of quadrats and there is no restriction to the number of quadrats. A smaller number of quadrats will increase the interaction among quadrats and in turn the CSR hypothesis may not be rejected, while a larger number of quadrats will result in the area of a quadrat too small to have enough points inside. Additionally, a better test was performed to assure the result.

The Kolmogorov-Smirnov test is based on the cumulative probability distribution, it compares the maximum difference between empirical distribution and theoretical distribution. In our test, the theoretical model is CSR. In the test, each data point was evaluated by a real-valued function T(x, y) which was defined at all locations in the window. Then the predicted distribution of the value of function T which is under CSR assumptions was compared with the observed distribution of the value of function T. We defined two function T to evaluate the observation data. We used x coordinate of each point as the function.

This test was conducted on all observations over the study area. Once the CSR test is performed, a suitable model for exploring the intensity of observation based on its spatial pattern can be considered.

3.2.2. Point process model

The primary goal of this research is to investigate the connections between the observation events and their related environmental conditions, to characterize the spatial variation of the occurrence of volunteered observations. Accordingly, point process models (PPMs) provide a framework for analysing the spatial structure of point patterns as the function of other spatial structures which can be the inherent characters of point object or regionalised spatial covariates. This method is used to answer research questions 3.1 and 3.2.

Based on the exploratory analysis of the spatial distribution of observations from partner groups, we found observations from the same partner group are spatially dispersal. So we assume observations collected by both partner groups and individual observers are independent of each other. Therefore, in this study, we focus on independent observations and the status of intensity. In the collection of volunteered phenological observations, some areas may be more attractive to volunteers for making observations. For instance, volunteers may prefer to make observations near where they live or places they are easily accessible. Besides, densely populated areas are more likely to acquire more observations. All

these potential factors are spatially varying and have an influence on data collection by volunteers. And the covariate information is defined at every point in the entire area. Thus, we formulated the model for the intensity function that incorporates the effects of these spatial covariates. An inhomogeneous Poisson process model is used.

In the model, we assumed a finite set y of points in an area u with the records of their locations. The intensity of the point process is the function λ(u) of a set of spatial covariates with the values of X(u) at every location in the area. The relationship is modelled as a realisation of an inhomogeneous Poisson process :

$$\lambda(u) = \rho(X(u)), \tag{3-2}$$

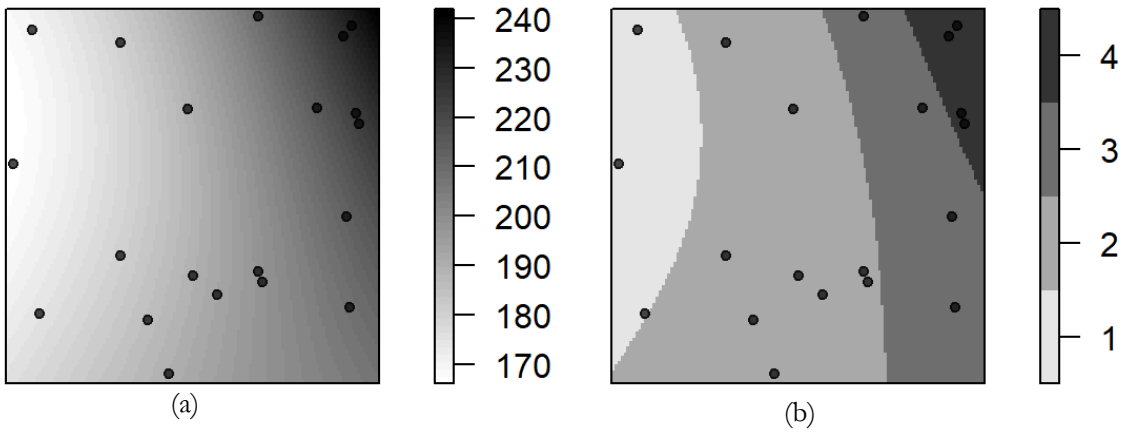
where the intensity function is dependent on covariate values X(u), and ϱ is the function of interest. We estimated the parameters using the package "spatstat" in R. The Berman-Turner algorithm is implemented in the package to fit the Poisson process model (Baddeley, 2010). The intensity function was modelled as a loglinear regression:

$$log\,\lambda(u) = \beta^T\big(X(u)\big), \tag{3-3}$$

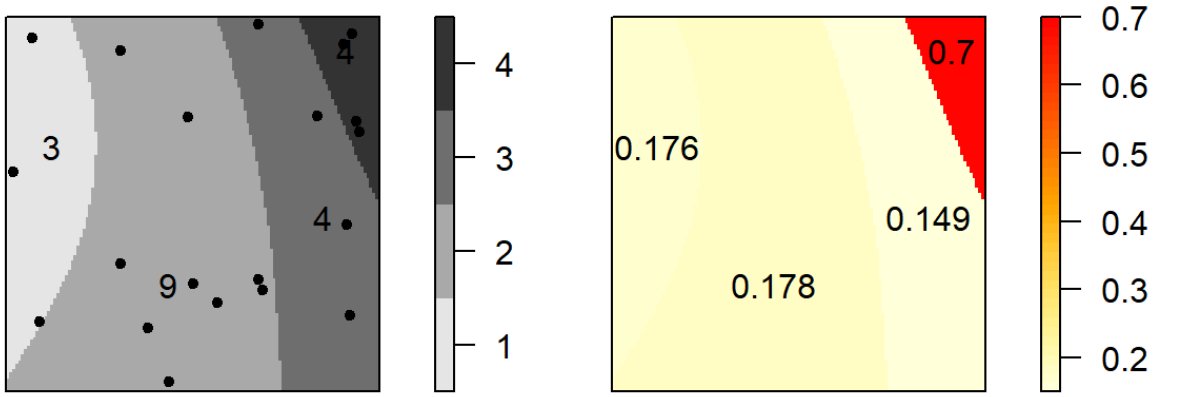
where β is a parameter vector, X(u) is a real-valued spatial covariate at every spatial location u.

Before estimating the effect of each covariate, the dependence of the observation pattern on each spatial covariates is explored. We used "local density" to visually show the dependence of the observation density on each covariate (Gimond, 2019). This approach helps to examine the variation of the underlying intensity across the study area. Also, it serves as an exploratory analysis to find out the potential relationship between covariates and observation density. The rationale of local density is:
(1) The entire study area is split into several sub-areas defined by equal range of covariate values;
(2) Observations within each sub-region are counted and calculated the intensity by dividing the number of observations in each quadrat by the area of the quadrat.

Figure 10 (a) – (d) show a detailed example that how an observation intensity map based on a covariate is produced. In the study, we split the covariate into four regions with equal interval of covariate value and calculated the intensity. This method can roughly find out whether the observation density prefers a particular range of covariate values.



(a) (b)



(c) (d)

Figure 10 An example of an intensity map based on a covariate (Gimond, 2019). (a) The elevation map with the distribution of point events. (b) The elevation value was first divided into four ranges with the same interval and shown on the map. Each sub-region was assigned a number from 1 to 4. (c) Counting the number of points in each sub-region (quadrat). (d) Calculating the point intensity within each sub-region (quadrat). The color bar shows the value of intensity.

Accordingly, we modelled volunteered observations at spatial locations as the intensity of observation, which is the function of a series of environmental conditions. That is, volunteers' preference in site selection leads to the variation in spatial coverage of observations. We fitted the model with seventeen covariates, and the Poisson point process is modelled with intensity:

$$\begin{aligned}\lambda(u) = exp\big(\alpha &+ \beta_1 GP(u) + \beta_2 RD(u) + \beta_3 LC1(u) + \beta_4 LC2(u) + \beta_5 LC3(u) \\ &+ \beta_6 LC4(u) + \beta_7 LC5(u) + \beta_8 LC6(u) + \beta_9 LC7(u) + \beta_{10} LC8(u) \\ &+ \beta_{11} LC9(u) + \beta_{12} LC10(u) + \beta_{13} LC11(u) + \beta_{14} LC12(u) \\ &+ \beta_{15} LC13(u) + \beta_{16} LC14(u) + \beta_{17} LC15(u)\big),\end{aligned} \tag{3-4}$$

where β_n denotes the effects of covariates that to be fitted: (1) gridded human population density GP, (2) gridded road density RD, (3) 15 different classes of land cover LC_n, α is the parameter to be fitted.

In practice, a PPM is fitted by maximum likelihood approach, which is written as:

$$L(\lambda) = \sum_{i=1}^{n} log\, \lambda(u_i) - \int_A \lambda(u) du, \tag{3-5}$$

$$\int_A \lambda(u) du \approx \sum_{i=1}^{m} \omega_i \lambda_i \tag{3-6}$$

where the integral indicates the expected number of points in the entire study area A. So to fit a PPM in the form of (3-4), variables should be estimated at every point in the entire study area A. However, our data only contains the presence data – collected observations, which do not cover the entire study area. To estimate the parameters of each covariate, a classical approach called "quadrature" is used to approximate the integral (Renner et al., 2015). In general, a set of "quadrature points" are generated in which the intensity function is estimated, then the integral is estimated by the sum of weights of quadrature points (3-5) (Warton & Shepherd, 2010). In the R package "spatstat", a set of dummy points that denotes the absence of points are automatically generated when fitting an inhomogeneous point process model.

For continuous data, the assessment of the goodness-of-fit of a fitted inhomogeneous PPM is done by analysing the residual defined for each observation. The definition of residual in a fitted PPM is the discrepancy between observed intensity and theoretical predicted intensity. Residual analysis can both

display heterogeneity and spatial trend in the data. The residual can be regarded as the variation under the effects of spatial covariates. It is written as (Baddeley, Turner, Møller, & Hazelton, 2005):

$$R(B) = n(\mathrm{x} \cap B) - \int_B \hat{\lambda}(u)du, \quad (3\text{-}7)$$

for any disjoint subregion B in the entire study region W, where x denotes the observed point intensity, $n(\mathrm{x} \cap B)$ denotes the number of points in B, $\hat{\lambda}(\mathrm{u})$ is the fitted intensity at all location u in W. The calculation of residuals not only includes points of the pattern but also attributes to the absence of points ("quadrature points"). We visualised the residuals by a diagnostic tool "diagnose.ppm" defined in "spatstat" to display the spatial trend of the residuals. However, this interpretation of residuals depicts the trend of cumulative residuals of each subregion B. Subsequently, a smoothed residual plot based on kernel smoothed estimation at location u was created. The smoothed residual presented as a contour form to display the trend of residuals. Figure 11 shows an example of a smoothed residual plot, where the residual ranges from-0.004 to 0.006. The 0 value denotes a good fit. The use of residual plot enables detecting spatial variation in the covariate setting.

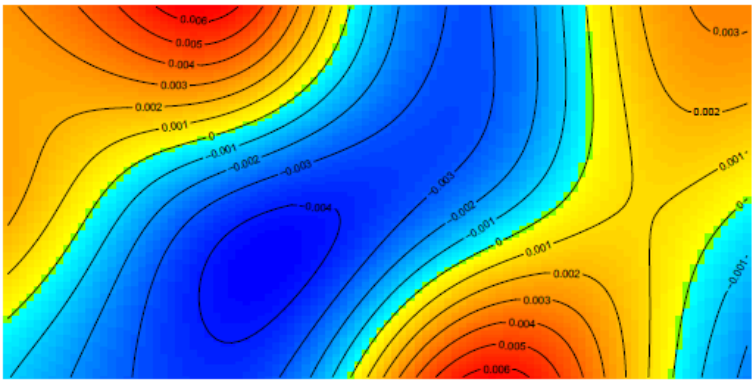
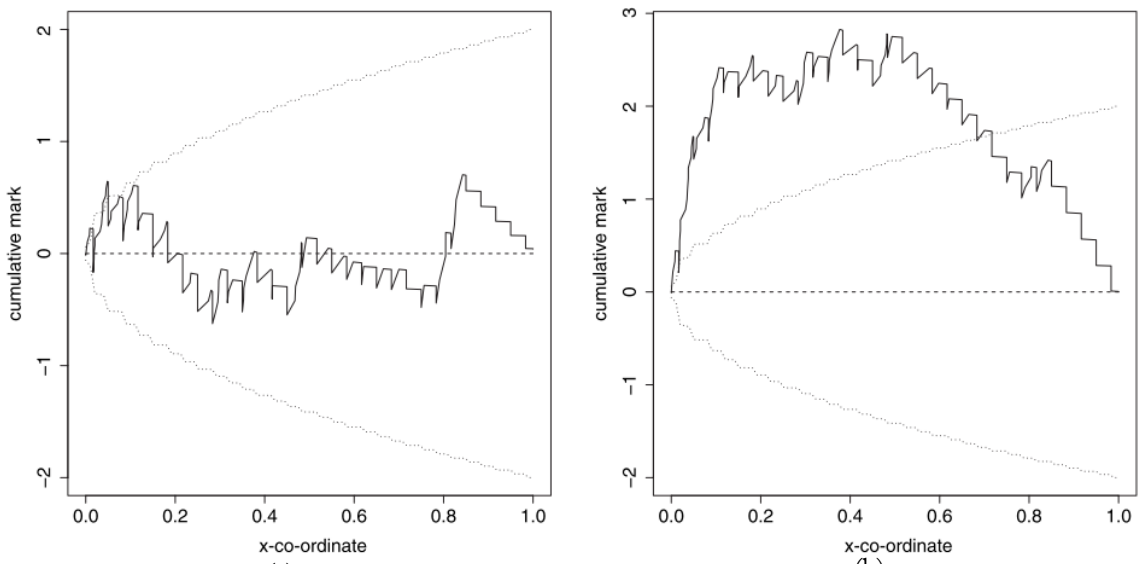


Figure 11 An example of a kernel smoothed residual field for a fitted heterogeneous Poisson model

Furthermore, we extracted the residual value of each point and created a residual map to find out the unusual patterns. The residual indicates the expected pattern after considering the effect of covariates. For example, a negative residual value indicates the model overestimates the intensity under the function of all spatial covariates.

In spatstat, the use of "diagnose.ppm" enables another test on the fitted spatial trend, which can help interpret the departure of the fitted distribution from the true distribution. It is obtained by plotting the derived residuals against a lurking variable. A lurking variable (which also called a confounding variable) is not included in the fitted model but influences the assessment of the relationship between the response variable and explanatory variables. Figure 12 displays two examples of a commonly used lurking variable in the residual analysis. The lurking variable is denoted by the x coordinates (Cartesian coordinate) of the point events in the defined study area. Plotting residuals against a lurking variable is to investigate the degree of deviation between the predicted pattern and the observed patterns. In the examples, the residuals are derived from a fitted inhomogeneous Poisson process model. The dotted envelope is the 95% confidence level for the cumulative raw residuals based on the variance under the model. In Figure 12 (a), except for a violation at a small x coordinate, most of the x fall into the dotted envelope, which indicates that the fitted spatial trend is consistent with the true spatial trend. Figure 12 (b) with a large violation indicates the spatial trend is not appropriate.



(a) (b)

Figure 12 Examples of lurking variable plots (Baddeley et al., 2005)

4. RESULTS

This section describes the results of applying the research methods. After data preparation, volunteered phenological observation data were organized into a set for subsequent model construction and validation.

The data filtering of the phenological dataset was implemented in Python 3.7, with libraries "pandas" and "numpy". The construction and the validation of the PPM were implemented in R 3.6.2. The library "spatstat" was used to build and validate the model, library "ggplot2" was used to create histograms. All the code and data can be found in https://github.com/XinyiQ/Investigating-spatial-bias-in-citizen-science-phenological-data.

4.1. Identification of spatial point patterns

The test of complete spatial randomness aims to answer the first research question and to find out whether the intensity of observations changes throughout the study area.

The selection of a number of quadrats played a significant role in the test result. To assure the reliability of the result, we considered 3 different forms of quadrat counts. According to equation 3-1, the study area should be divided into user-defined quadrats. We tried several sets of divisions, but it encountered a problem that the quadrats are too small to have sufficient point events inside when dividing more than 400 grids, which would bring uncertainty to the estimation of the results. This is due to the insufficient number of observations across the very large study area, and thus not sufficient observations can be divided into some quadrats to implement the estimation. After experimenting, the area was planned to divide into 150 grids, 64 grids, and 30 grids. But due to the irregular shape of the study area, it was actually divided into 107 grids, 57 grids, and 21 grids. Table 4.1 shows the combination of quadrats and the associated hypothesis testing results based on Pearson χ^2 test.

We defined the significant level as 0.001 because the frequency of observation events in a subregion is quite small and the smaller the p-value is, the less likely under the null hypothesis to be rejected.

Table 3 The results of quadrat counts test

Quadrat counts m	χ^2_{m-1}	P-value
10 * 15	11294	< 0.001
8 * 8	7231.8	< 0.001
5 * 6	5873.5	< 0.001

The P-value for all different sets of quadrat counts provided strong evidence to reject the uniform Poisson process. The large χ^2values indicate a great departure of the target point pattern from CSR. We can, therefore, assume the point pattern follow a non-uniform distribution, or the point events are not independent of each other. But in our case, we assumed the observations across the dataset have no interaction with each other. Thus, the point pattern of volunteered phenological observations follows an inhomogeneous Poisson process.

Considering the limitation of CSR, and in order to assure the reliability of the test result, we used another test Kolmogorov-Smirnov to evaluate CSR. Figure 13 presents the results of the K-S test under function T(x, y). We used the x coordinate of the point in the Cartesian coordinate system to set as the function T

to perform a goodness-of-fit test. The red dotted line denotes the expected distribution which represents the uniform distribution, and the black line denotes the observed distribution which represents the real distribution of all point events. The x-axis represents the x Cartesian coordinate of point events. Several bearings around the x Cartesian coordinate of 500000, 1100000, 1400000, 1800000 indicate some aggregation of points within the entire area and a non-homogeneous distribution to some extent. From the x value of 1200000, it experienced a relatively significant rise, which indicates there are more aggregated point events at the larger x Cartesian coordinates, which is the eastern part of the real region. Besides, the large difference between the two lines indicates the departure of the real distribution from the uniform distribution. With the low of the p-value, we can suggest the rejection of the assumption of CSR.

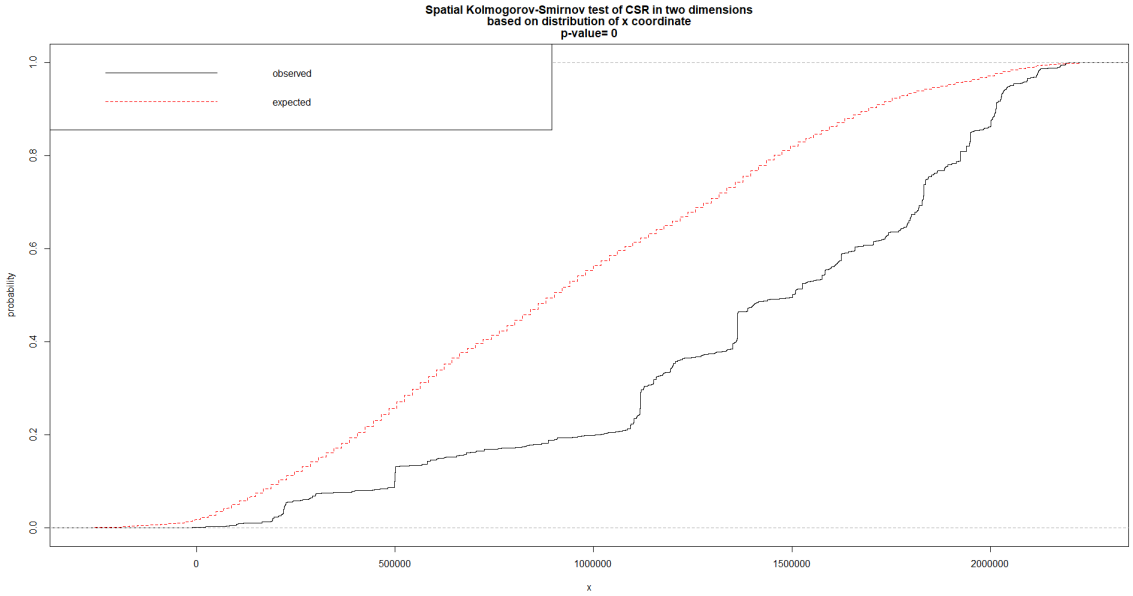


Figure 13 Result of Kolmogorov-Smirnov test for Complete Spatial Randomness

Figure 14 shows the spatial distribution of observations and Figure 15 shows the land cover map of the entire study area. The point pattern of observation shows an evident tendency that there are more observations in east than in west. There are several visible clusters near the northeast region. Observations in the southwest region are sparse and spatially dispersed. Figure 16 (a)-(q) show the spatial distribution of each spatial covariate, which helps better understand the potential relationship between variables. According to the population density, the study area was divided into four parts with an equal interval of density values (Figure 16 (a)). However, the difference in population density between regions is not large, except for a few cities which are not evident shown in the map. Likewise, only a few areas have relatively high road density in the map. Moreover, combined with the distribution map of observations, there is no obvious spatial trend in areas with high road density, and it does not reflect the potential spatial relations with the observation intensity. What's more, the coverage of different land cover classes varies largely in space. The original land cover map was reclassed to 15 land cover classes. It was then calculated the coverage of each class in a defined unit. The coverage was then calculated by counting the area of each class within a defined grid cell. The study area was also split into four parts based on the same interval of coverage. Deciduous forest has the most coverage, but the densest forest areas (central region) do not show a higher number of observations. In the northeast part, the coverage of the mixed forest is higher, and at the same time observations are spatially aggregated. Areas dominated by evergreen forest have very few observations and is sparsely distributed. Areas covered by cultivated crops near the northwest central region also have several visible clusters.

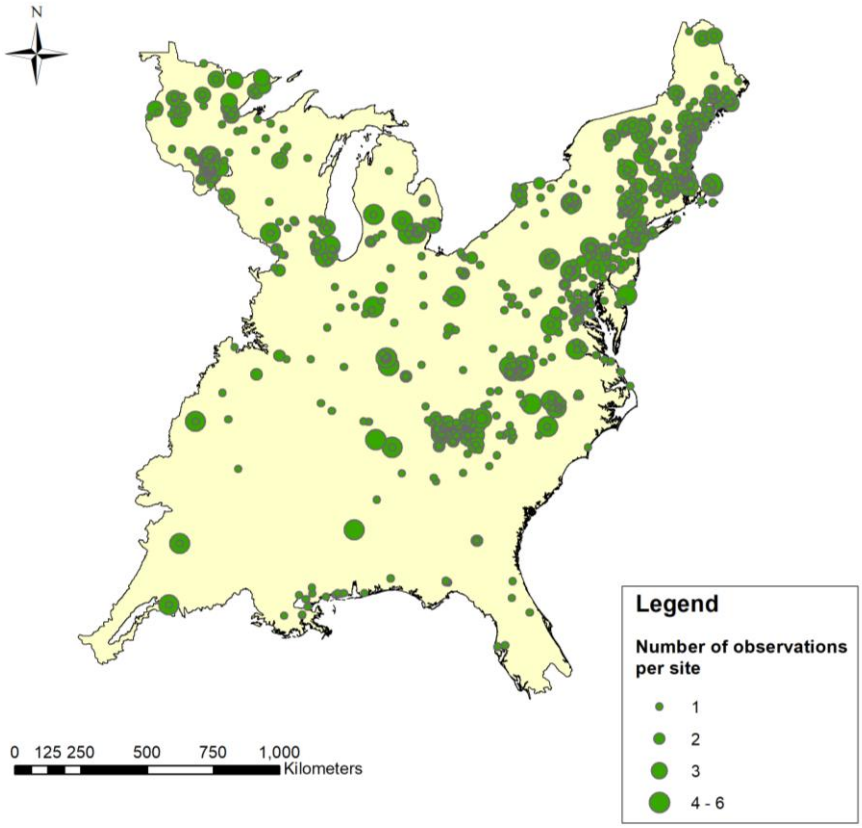


Figure 14 The spatial distribution of volunteered phenological observations

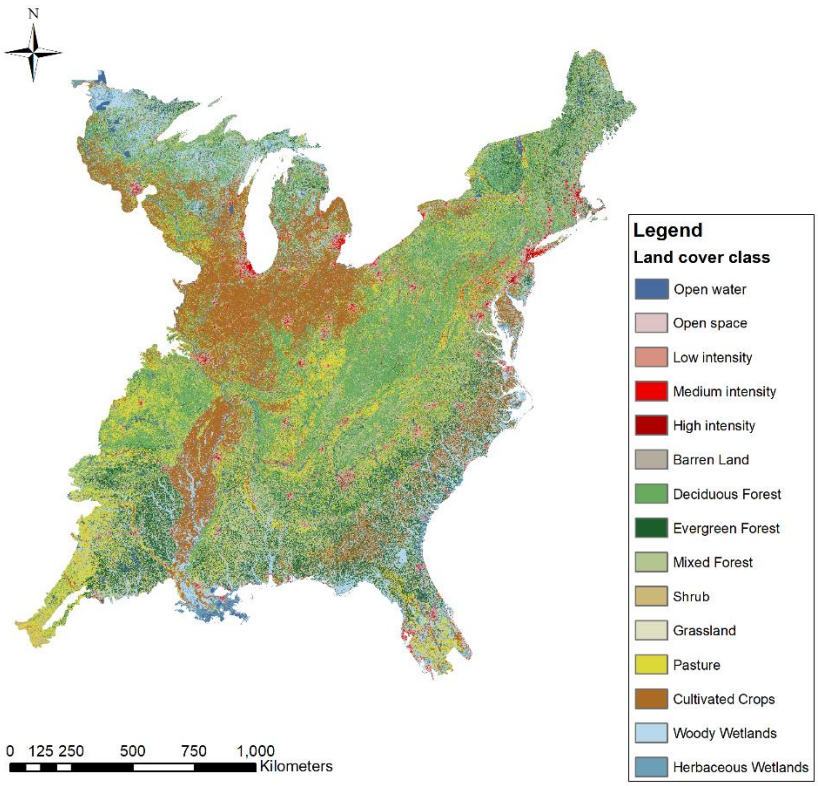
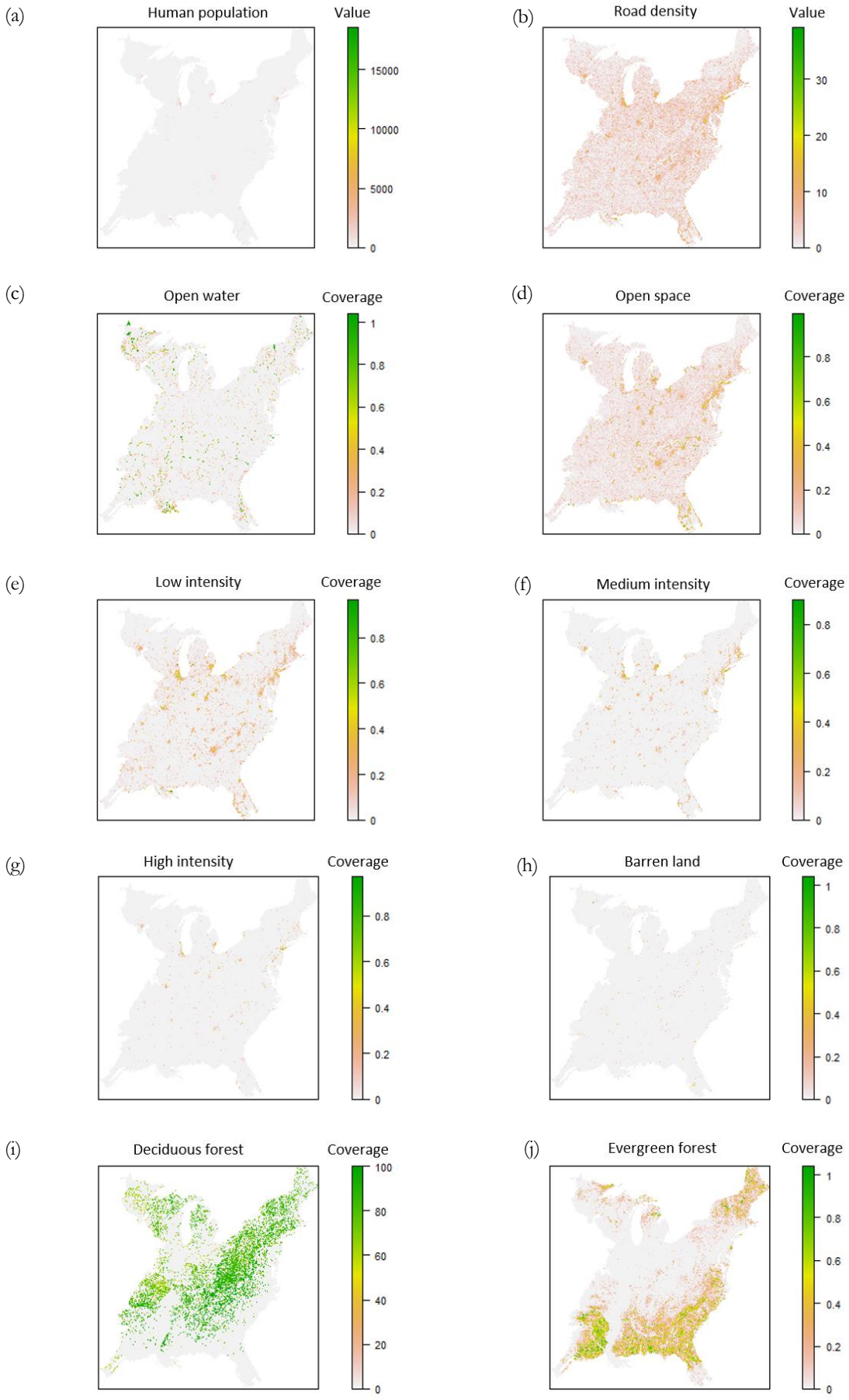


Figure 15 Land cover map


(a)
Human population
Value
15000
10000
5000
0
(b)
Road density
Value
30
20
10
0
(c)
Open water
Coverage
1
0.8
0.6
0.4
0.2
0
(d)
Open space
Coverage
0.8
0.6
0.4
0.2
0
(e)
Low intensity
Coverage
0.8
0.6
0.4
0.2
0
(f)
Medium intensity
Coverage
0.8
0.6
0.4
0.2
0
(g)
High intensity
Coverage
0.8
0.6
0.4
0.2
0
(h)
Barren land
Coverage
1
0.8
0.6
0.4
0.2
0
(i)
Deciduous forest
Coverage
100
80
60
40
20
0
(j)
Evergreen forest
Coverage
1
0.8
0.6
0.4
0.2
0

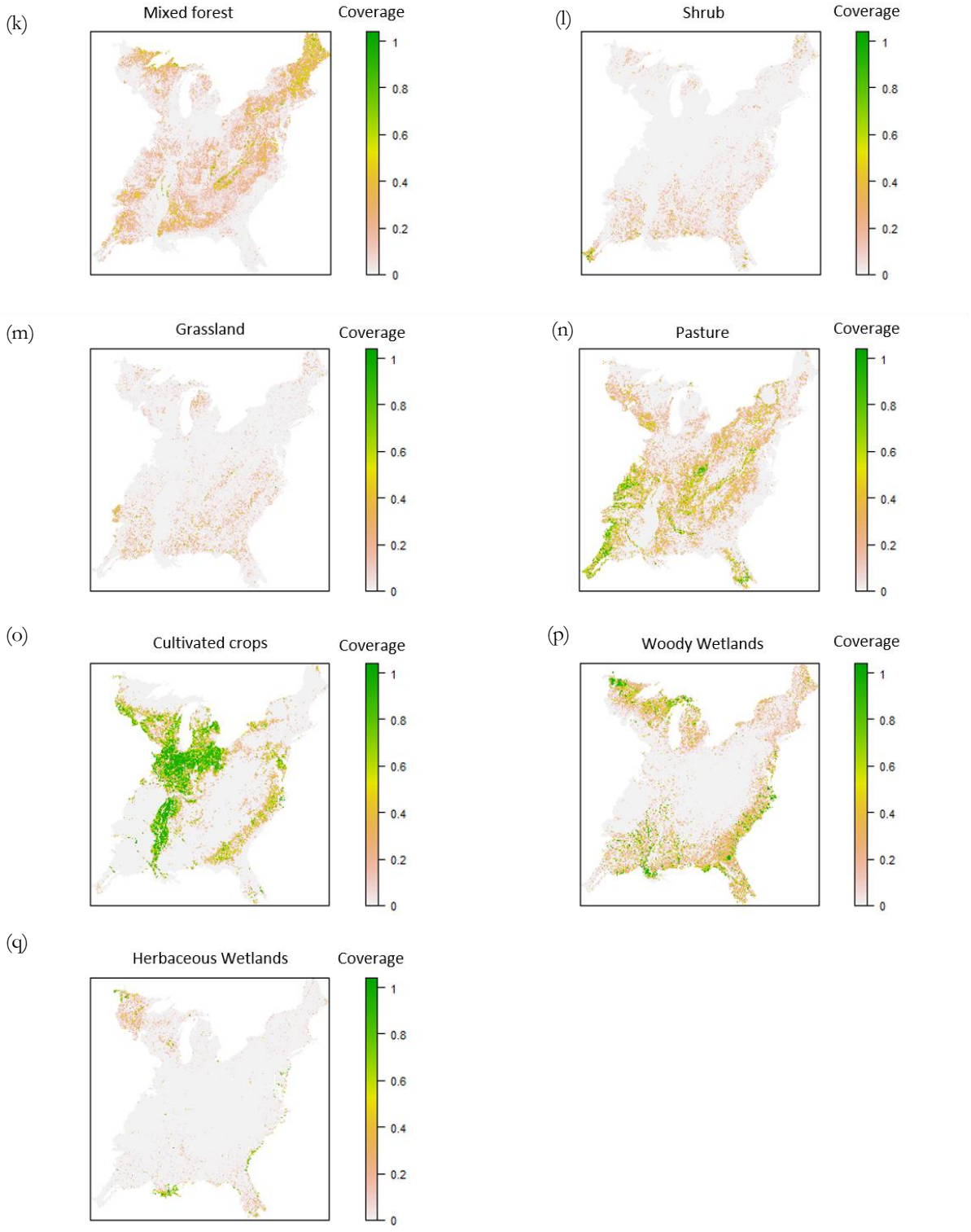


Figure 16 The spatial distribution of each covariate. The "coverage" means the share of each land cover within a grid cell. Because we calculated the area of one specific

4.2. Identification of spatial bias

The exploratory analysis of dependence between covariates and point density can bring insights into identifying potential spatial bias. It is of interest whether the intensity depends on specific spatial covariates, as is shown in Figure 17 (a) – (q). The color bar indicates the observation intensity within four defined quadrat regions. In general, red indicates areas with higher observation intensity. Some of the maps do not display the defined four sub-regions and only two sub-regions are visible, because the cover of those sub-regions is so small that they are not clearly shown on the map.

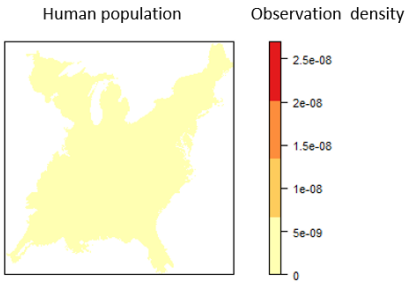
Due to the large extent of the study area, it failed to show the relationship between human population density and observation density. In fact, several small regions with high observation density were found only if the figure is large enough, so they are not visible in Figure 17 (a). In Figure 17 (b), red pixels represent areas with both high observation density and high road density, which shows a strong dependence between these two variables.

In Figure 17 (c) – (q), there are several land cover types show very strong dependence. For instance, red open space areas (Figure 17 (d)) tend to have higher observation intensity. This is because the coverage of open space regions is small, when compared with the other three sub-regions with a larger area, even a small number of observations will result in a larger observation intensity. Other land cover classes have a similar tendency, such as medium intensity, low intensity, high intensity, barren land, and deciduous forest. Particularly, deciduous forest areas with high observed intensity (Figure 17 (i)) are not as complete as their coverage in the entire study area (Figure 16 (i)). This is because only a part of the observations is in areas with deciduous trees.

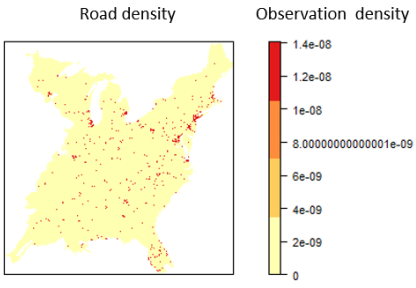
Figure 17 (k) and (q) have a similar tendency, where areas with the specific land cover type (mixed forest & herbaceous wetlands) have higher observation intensity, with the rest areas with relatively lower observation intensity.

In Figure 17 (c), most areas are covered with red, which means areas without this specific land cover type (open water) have a higher observation intensity. Similarly, low intensity, evergreen forest, shrub, grassland, pasture, cultivated crops, woody wetlands have the same tendency.

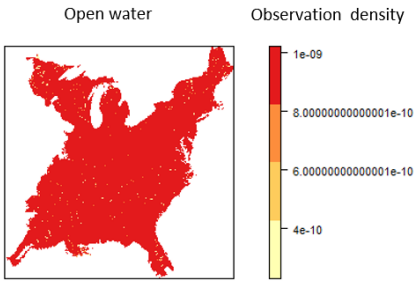
(a)


Human population
Observation density
2.5e-08
2e-08
1.5e-08
1e-08
5e-09
0

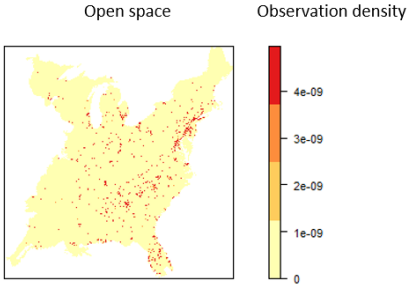
(b)


Road density
Observation density
1.4e-08
1.2e-08
1e-08
8.00000000000001e-09
6e-09
4e-09
2e-09
0

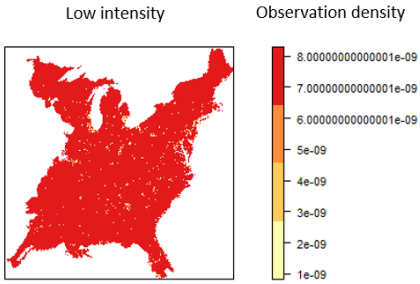
(c)


Open water
Observation density
1e-09
8.00000000000001e-10
6.00000000000001e-10
4e-10

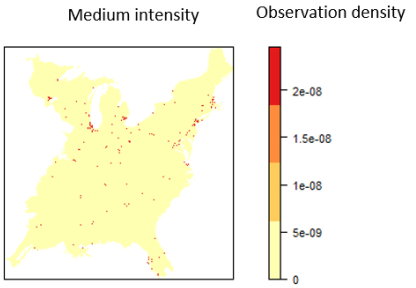
(d)


Open space
Observation density
4e-09
3e-09
2e-09
1e-09
0

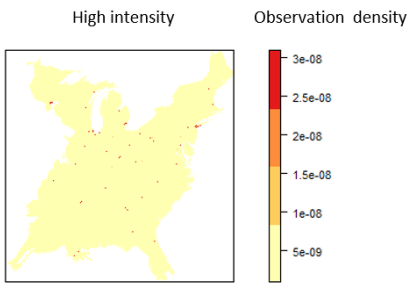
(e)


Low intensity
Observation density
8.00000000000001e-09
7.00000000000001e-09
6.00000000000001e-09
5e-09
4e-09
3e-09
2e-09
1e-09

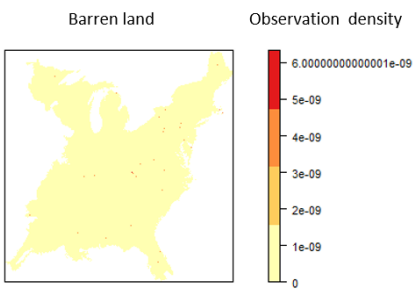
(f)


Medium intensity
Observation density
2e-08
1.5e-08
1e-08
5e-09
0

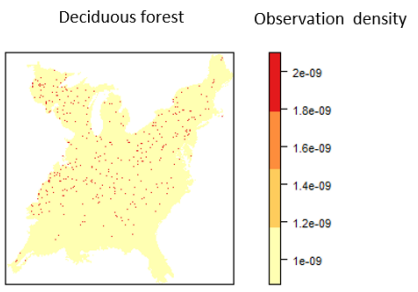
(g)


High intensity
Observation density
3e-08
2.5e-08
2e-08
1.5e-08
1e-08
5e-09

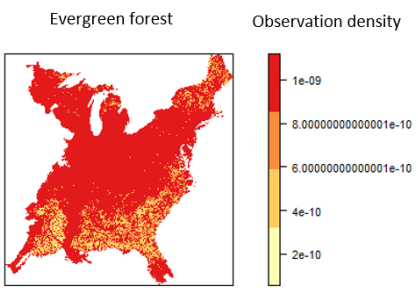
(h)


Barren land
Observation density
6.00000000000001e-09
5e-09
4e-09
3e-09
2e-09
1e-09
0

(i)


Deciduous forest
Observation density
2e-09
1.8e-09
1.6e-09
1.4e-09
1.2e-09
1e-09

(j)


Evergreen forest
Observation density
1e-09
8.00000000000001e-10
6.00000000000001e-10
4e-10
2e-10

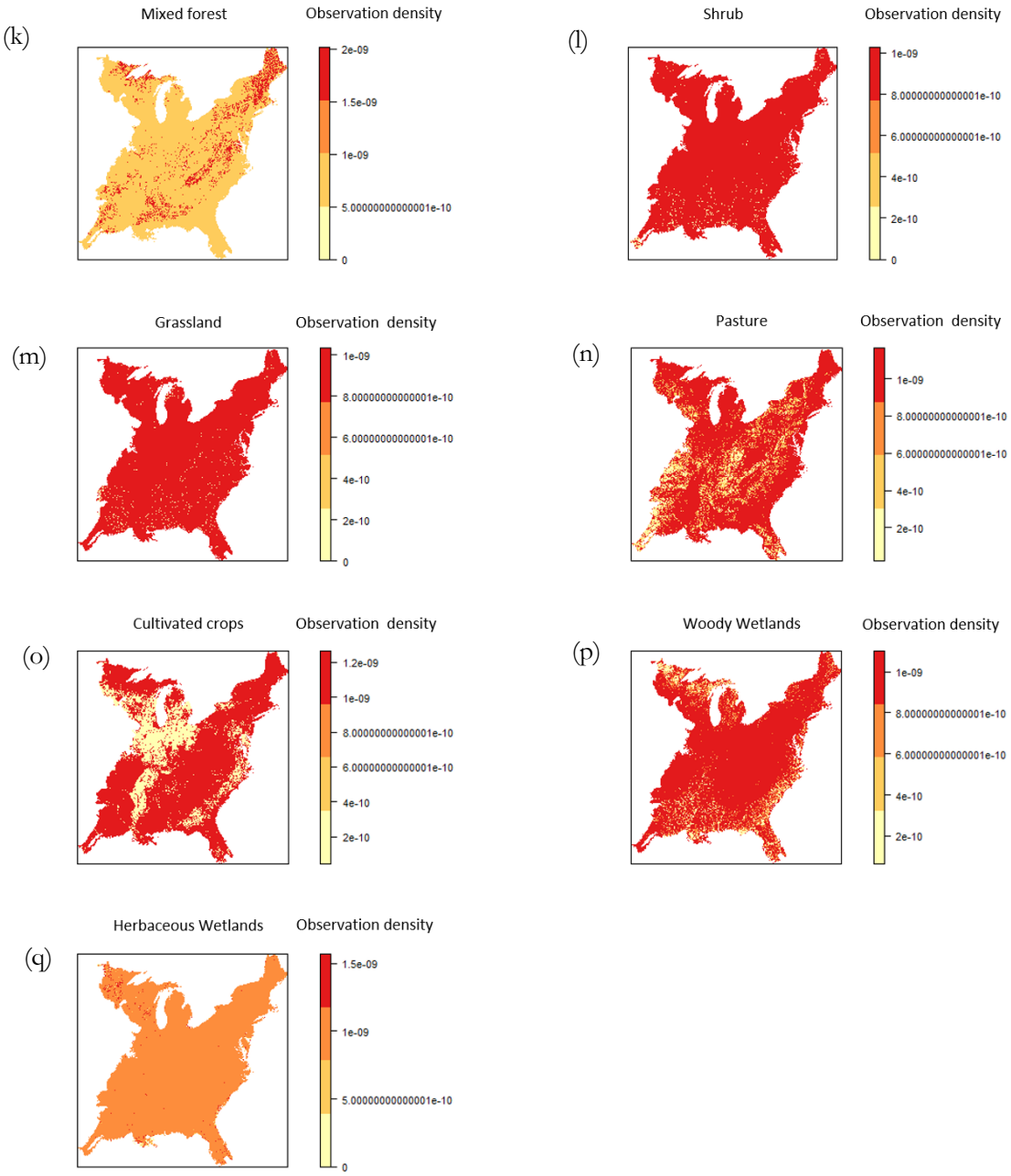


Figure 17 Local density map of observation intensity on each covariate

To model the effect of our proposed spatial factors on sampling effort, we fitted an inhomogeneous Poisson process model. To fit the model, 5652 dummy points are not observed points in the pattern generated in space. In spatstat, the dummy points are generated in a regular grid and the distribution is shown in Figure 18 (a). For better visual inspection, we extracted a subarea. Combined with the sample observation points, there are in total 8568 points in the entire study area and the distribution is shown in Figure 18 (b), dummy points were shown in the form of dots. The circles indicate the observation point, and the grey scale of circles is proportional to the number of observations.

(a) (b)

Figure 18 The distribution of the point pattern

The initial analysis of the point pattern is fitting the inhomogeneous Poisson process model. The best-fitted model was selected based on the Akaike Information Criterion (AIC) using the automatic stepwise deletion. The stepwise deletion measures each term in the full model and optimally deletes the term to obtain the best-fitted performance and the combination of terms. In our case, the initial model contains all covariates, but then only one different combination of covariate was automatically measured. The initial selection of the model is shown below.

$$\begin{aligned} log\,\lambda(x) = \alpha &+ \beta_1 pop + \beta_2 rd + \beta_3 lc11 + \beta_4 lc21 + \beta_5 lc22 + \beta_6 lc23 + \beta_7 lc24 + \beta_8 lc31 + \beta_9 lc41 \\ &+ \beta_{10} lc42 + \beta_{11} lc43 + \beta_{12} lc52 + \beta_{13} lc71 + \beta_{14} lc81 + \beta_{15} lc82 + \beta_{16} lc90 \\ &+ \beta_{17} lc95 \end{aligned} \tag{4-1}$$

The parameter estimation is based on a loglinear regression. β_n represents the estimated effect of covariates: human population density "pop", road density "rd", 14 different land cover classes "lc_n". The AIC value of model (4-1) equals to 120145.9.

With the automatic stepwise deletion provided by the spatstat package, another model was selected, which is stated below.

$$\begin{aligned} log\,\lambda(x) = \alpha &+ \beta_1 pop + \beta_2 lc11 + \beta_3 lc21 + \beta_4 lc22 + \beta_5 lc23 + \beta_6 lc24 + \beta_7 lc31 + \beta_8 lc41 \\ &+ \beta_9 lc42 + \beta_{10} lc43 + \beta_{11} lc52 + \beta_{12} lc71 + \beta_{13} lc81 + \beta_{14} lc82 + \beta_{15} lc90 \\ &+ \beta_{16} lc95 \end{aligned} \tag{4-2}$$

This model eliminates the function of road density, which improves the model AIC to 120144.7, an improvement of around 1 unit. Both of these two models were tested to compare the performance of models. However, there is not much difference in the result of parameter estimation. Therefore, the results discussed following are based on model 4-1.

The goodness-of-fit of the model was checked by interpreting the smoothed residuals from the fitted model (Figure 19 & Figure 20). The raw residual represents the difference between the observed intensity and the estimated intensity ranges from -4e-10 to 1.2e-09. The residual value indicates the difference between the true observation intensity and predicted observation intensity. In PPM, negative values indicate the model overestimates the intensity hence predict more observations than true values. Similarly, a positive value indicates that the observations have a higher degree of aggregation at that location. From Figure 19, around 66% of all residual values are negative. That is to say, under the influence of all spatial covariates, around 66% of observations are deemed as under-sampling. Especially, from a spatial perspective, the right bottom of Figure 20 shows the distribution of the estimated raw residuals. We found the north-western and east-central regions have the best fitted, which represents a proper intensity of observations. While the intensity of observation in most southwest region is relatively low. The north-eastern boundary of the study region has an apparently elevated observation intensity.

Figure 20 is a standard four-panel plot of the spatial trend of the estimated raw residuals. The upper left plot shows the data points (positive residuals) with circles and the fitted intensity (negative residuals) with the background color. The lower right plot shows the kernel smoothed residual field. The lower left and upper right show the fitted residuals against a lurking variable, where the cumulative residuals were summated from west-east and south-north directions. The dotted envelope is the 95% confidence level for the cumulative raw residuals based on the variance under the model. The x Cartesian coordinates and y Cartesian coordinates that do not fall into the 2σ-limits boundary indicate that the true spatial trend deviates from the fitted spatial trend.

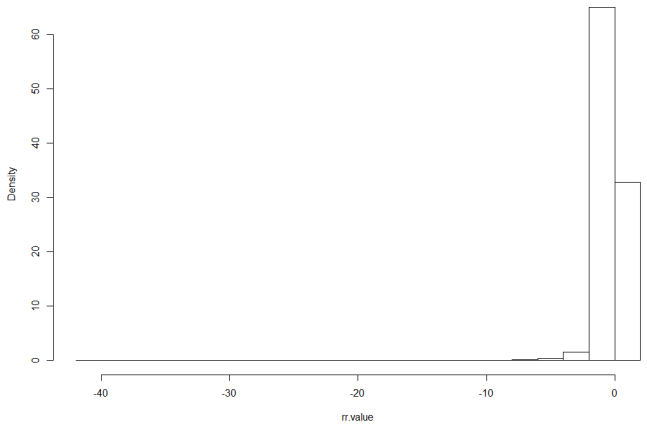


Figure 19 The frequency of raw residual from the fitted model

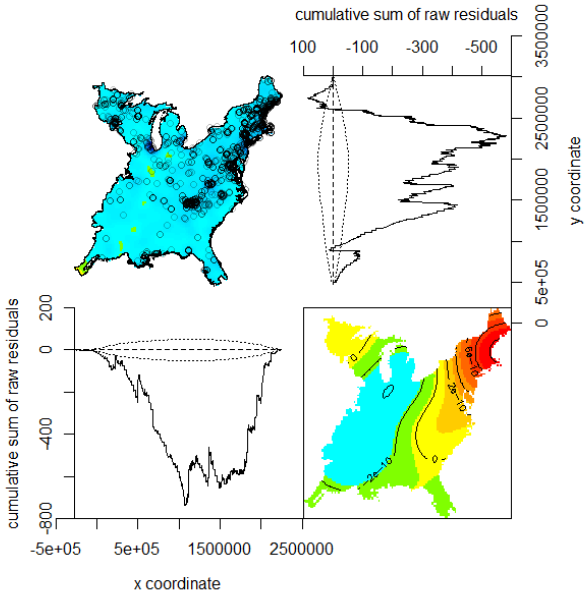


Figure 20 Four-panel plot of the estimated raw residuals

Figure 21 shows the fitted conditional intensity at two different angles. Potential observation points were generated at new locations according to the function of a set of covariates. We computed the expected number of observations in a grid cell with an area of 100 * 100 m. The length of the bar indicates the expected number of observations in an area. There is a peak intensity in the northeast region. The difference in observation intensity in other areas is not significant. Generally, there are several distinct small peaks in the central, western, and south-eastern regions. The overall expected observations are relatively sparse with respect to the broader spatial extent of our study area.

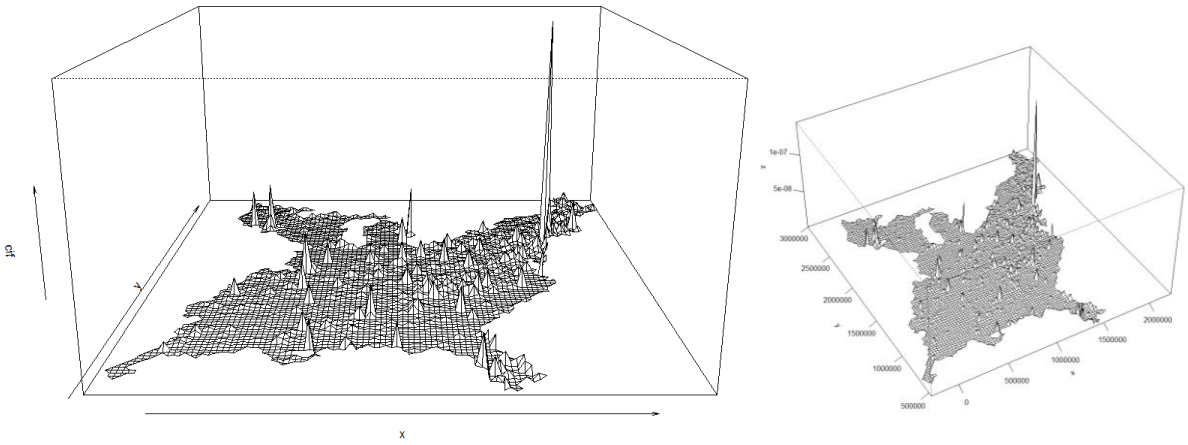


Figure 21 The predicted intensity of the point pattern

After fitting the log-linear regression model, the estimated parameters and statistical significance for all spatial covariates are shown in Figure 22 and Table 4. Human developed areas have different influences on the probability of observation, with the lower intensive developed area negatively affecting the intensity, with the medium intensitive developed area having the biggest effect. As expected, deciduous forest and mixed forest positively affect observation intensity. Mixed forest with low coverage of deciduous has the biggest effect size. And evergreen tree species have a negative impact which makes sense, as the target species is deciduous forest. In particular, there is a greater possibility of declining

observation intensity in the planted areas where crops are grown. Unexpectedly, barren has a significant positive impact on increasing intensity. Moreover, both road density and human population density have a significant influence.

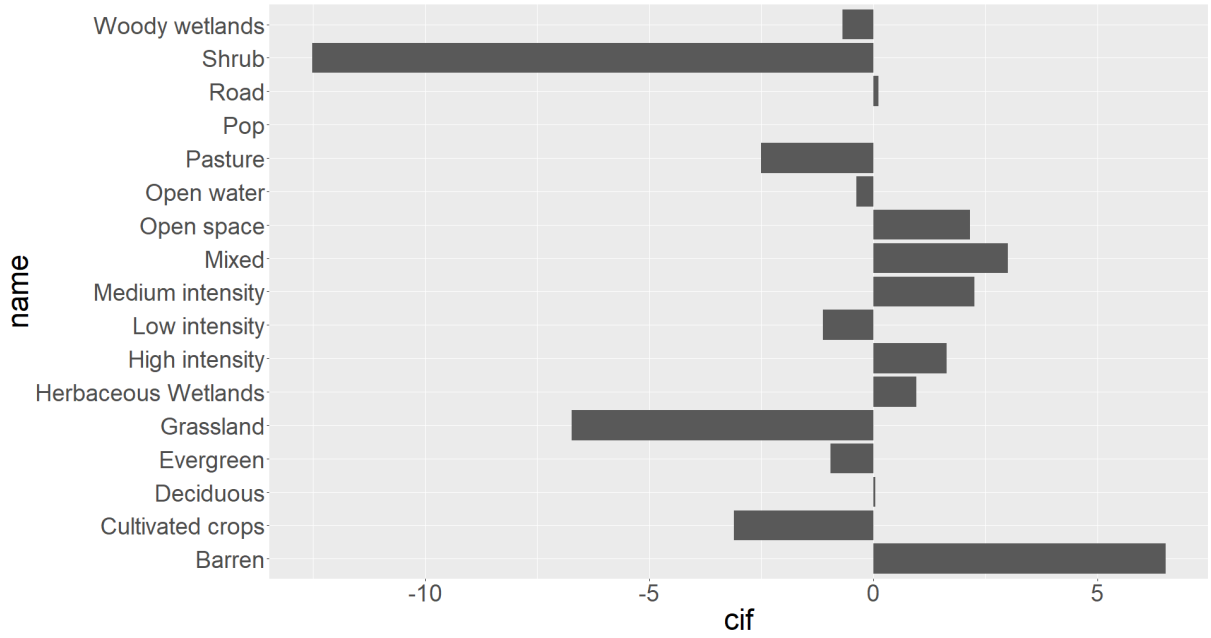


Figure 22 The effect size of all spatial covariates, where cif means the model is fitted at required location u.

Table 4 Estimated parameters of the fitted model for all spatial covariates

	Estimate	S.E.	CI95.lo	CI95.hi	Ztest	Zval
Intercept	-2.104849e+01	8.518197e-02	-2.121545e+01	-2.088154e+01	***	-247.1003140
Pop	4.765439e-04	1.986452e-05	4.376101e-04	5.154776e-04	***	23.9896951
Road	1.150109e-01	5.970409e-03	1.033091e-01	1.267127e-01	***	19.2634862
Open water	-3.693379e-01	2.194200e-01	-7.993931e-01	6.071740e-02		-1.6832462
Open space	2.166428e+00	1.876988e-01	1.798546e+00	2.534311e+00	***	11.5420487
Low	-1.119095e+00	2.098879e-01	-1.530468e+00	-7.077224e-01	***	-5.3318698
Medium	2.259195e+00	2.410409e-01	1.786764e+00	2.731627e+00	***	9.3726621
High	1.632196e+00	2.888962e-01	1.065969e+00	2.198422e+00	***	5.6497643
Barren	6.520606e+00	3.388385e-01	5.856495e+00	7.184717e+00	***	19.2439948
Deciduous	4.009321e-02	7.245882e-02	-1.019235e-01	1.821099e-01		0.5533241
Evergreen	-9.570481e-01	1.779851e-01	-1.305892e+00	-6.082038e-01	***	-5.3771257
Mixed	3.000203e+00	1.481070e-01	2.709919e+00	3.290488e+00	***	20.2570018
Shrub	-1.252378e+01	1.258481e+00	-1.499036e+01	-1.005720e+01	***	-9.9515011
Grassland	-6.735484e+00	7.965581e-01	-8.296709e+00	-5.174258e+00	***	-8.4557339
Pasture	-2.499295e+00	1.999533e-01	-2.891196e+00	-2.107393e+00	***	-12.4993933
Crops	-3.113423e+00	2.079462e-01	-3.520990e+00	-2.705856e+00	***	-14.9722509
Woody	-6.849068e-01	1.561548e-01	-9.909645e-01	-3.788491e-01	***	-4.3860767
Herbaceous	9.589067e-01	2.166794e-01	5.342230e-01	1.383590e+00	***	4.4254640

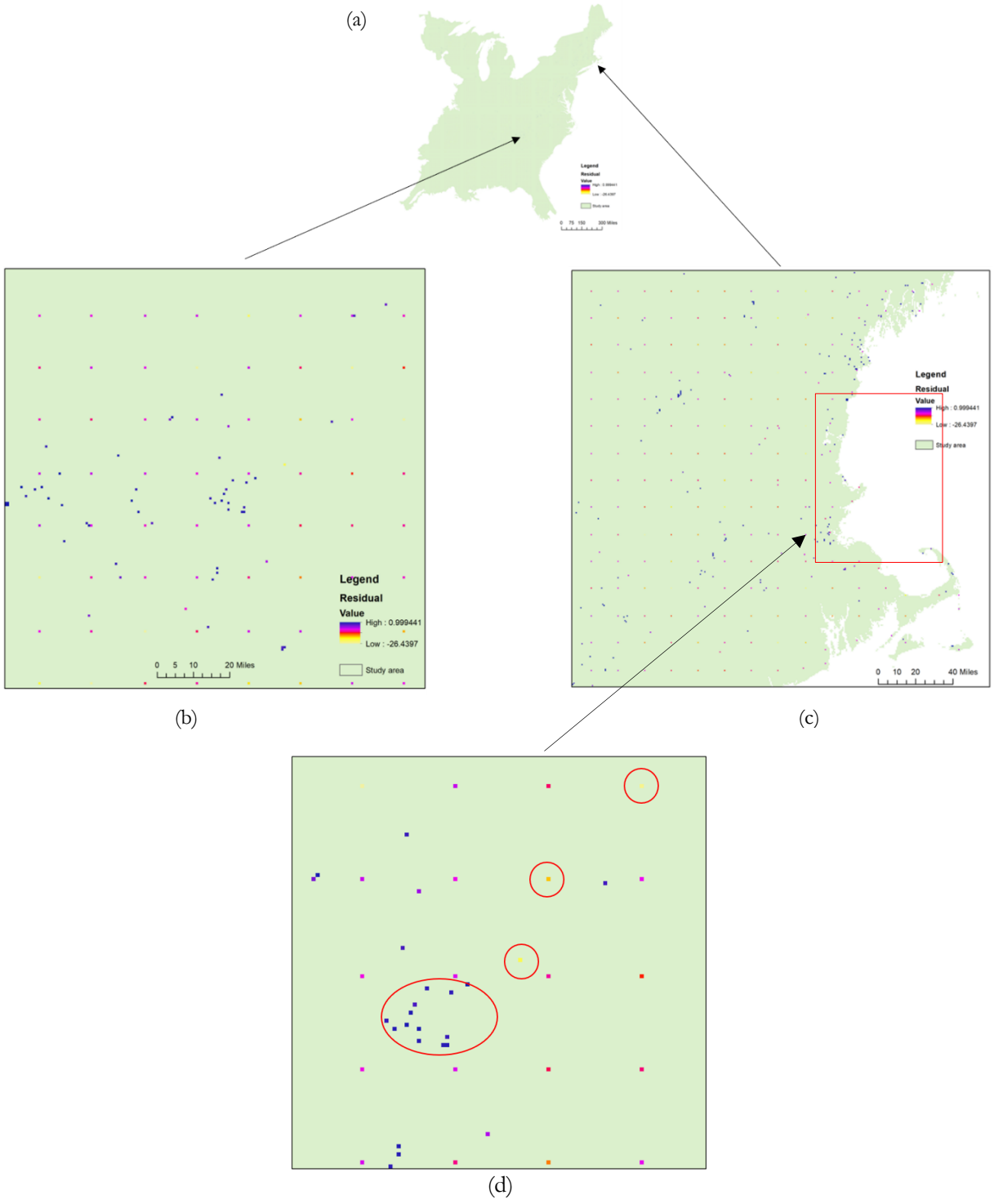


Figure 23 The residual maps which show the overestimation and underestimation of the fitted model; (a) The entire residual map; (b) & (c) Some clusters with positive residuals; (d) The residual value of three highlighted point from top to bottom are -2.497299, -1.491180, -2.087358, which means that the predicted intensity is higher than the true intensity by -2.497299, -1.491180, and -2.087358, respectively.

The geographic display of raw residual makes it easy to detect unusual patterns. It helps identify the systematic deviation of the individual grid cell containing observations from the predicted pattern. Figure 23 (a) shows an overall residual map of the observation. Due to the large scale of study area, each pixel that contains the residual value is not visibly interpreted on the map. Because the study area contains more

than 6 million pixels. Figure 23 (b) and (c) show several blue clusters, suggesting more observations than expected. Three points highlighted in Figure 23 (d) with large negative values suggest the inadequate sampling effort hence the lower-than-expected observation intensity clustered at that location. However, it is worth noting that most real observations in Figure 23 (d) are blue, representing overestimation. While most dummy points represent underestimation.

5. DISCUSSION

This chapter first summarises the answers to research questions. Secondly, interpretations and implications of results are presented. Also, limitations in the data materials and method as well as recommendations based on current results are discussed.

5.1. Answers to research questions

In this study, we examined the spatial bias in the distribution of observations arisen from surrounding environmental and social-economic factors during volunteered data collection and quantified the effects of the causes of spatial bias using an inhomogeneous Poisson point process model. The answers for specific research question are listed as follows:

Research question 1.1 What is the spatial pattern of volunteered phenological observations?
The spatial pattern is identified by testing the departure of an observed pattern from the complete spatial randomness (which also called homogeneous Poisson process) (Mateu, Usó, & Montes, 1998). The result of the CSR test shows the spatial pattern of volunteered observations is a non-uniform distribution. Meanwhile, the test of CSR is a prerequisite for the subsequent selection of a series of models. Two properties (intensity and inter-point interaction) of the Poisson point process bring different angles to understand the cause of a point pattern. In this study, we considered the property – intensity, which focuses on the intensity function raised by covariates. The distribution of observations is caused by volunteers' behavior in collecting data may influence the spatial pattern of observations.

Research question 1.2 What is the spatial bias in volunteered phenological observations?
The spatial bias in this study is defined as the uneven distribution of observations in a defined area. For phenology, the ideal experimental design would be a uniform coverage of observations over the study area both in spatial and temporal dimensions (Brunsdon & Comber, 2012). Brunsdon and Comber found the fluctuations in the onset date of plants have a significant relationship with the number of observations. Miller-Rushing et al. (2008) also pointed out the number of observations in an area have an impact on the variation in phenological records. To mitigate such effect caused by spatial bias on the construction of further phenological models, it is necessary to take spatial bias in the distribution of volunteered phenological observations into consideration.

Research question 2.1 Which spatial factors affect the spatial bias of volunteered observations?
The nature of the heterogeneity of voluntary data collection brings biases into records data. For one thing, volunteers may have a different preference in site selection. Volunteers may tend to collect observations near where they live, which makes areas with residential landscapes oversampled (Isaac & Pocock, 2015). For another, areas with higher human population density have a significant influence on sampling effort (Mair & Ruete, 2016). What's more, the higher spatial coverage of target objects or better accessibility can also attract more volunteers to make observations (Romo, García-Barros, & Lobo, 2006). Based on the literature review, in this study, we examined the spatial variation of volunteered observations that incorporates the effects of the human population, road density, and various classes of land cover.

Research question 3.1 How to identify the influence of these factors on the variation of the spatial distribution of volunteered observations?
We first visualized the distribution of observations and each spatial covariate (Figure 16). These plots show the potential relationship between observation clusters and covariate values. Further, the density

analysis based on each covariate brings an evident visual interpretation of potential bias in sampling effort (Figure 17).

Research question 3.2 To what extent these factors explain spatial bias?
We estimated the effect size of covariates representing spatial bias. As is expected, the increase in the human population translates into an increase in the number of observations. Similarly, more number of participants prefer to collect observations from high road accessibility areas. Such results demonstrate the importance of these variables in affecting sampling effort (Mair & Ruete, 2016). However, in contrast to road density, the impact of the human population is relatively small. This is likely because the gridded population does not vary much in space, except for a few extremely densely populated cities.

What's more, we found that the tendency of observation intensity varies according to different vegetation and land use types. Particularly, the three forest types have both different sizes and degrees of effects. The evergreen forest has a negative effect which makes sense because the main species to be collected is deciduous trees. The mixed forest which is dominated by both deciduous and evergreen species significantly affects sampling effort. However, the effect size of the deciduous forest is not significant compared with other types of land cover. We assume this result could be caused by two aspects. First, the overall coverage of deciduous coverage is largely more than two other forest types (Figure 16 – (i)). Moreover, there are quite large areas covered by deciduous forests that lack a sufficient number of observations. The deciduous tree forest area with higher observation density only covers a small part of the study area, and most of the deciduous forest region has a relatively smaller observation intensity (Figure 17 – (i)). And it can partly explain the relatively small effect size of deciduous forest – with the estimated coefficients 0.04. Second, volunteers are free to select sites for making observations (Koch et al., 2015). Thus, it is likely that observation may occur anywhere as long as the target species is present, not necessarily in deciduous forest areas. For example, observers may record observations at their back yard considering the convenience of daily recording. Likewise, observers may make observations along the roadside or streets occasionally. Also, the estimated accuracy of the land cover classification is about 84% (Fry et al., 2011), which may bring uncertainty in determining the actual land cover in a particular area. As for other land cover types, we found barren land has a strong and relatively large impact on controlling the changes in observation intensity. According to the visual interpretation of observation density with barren land (Figure 16 – (h), areas with less cover of barren land tend to have more observations. But in accordance with the density map (Figure 17 – (h), volunteers have a higher observation intensity in barren land regions. We also found this tendency among other land cover types which have less coverage over the study area but have bigger effect size, such as shrub, grassland, and pasture. This could be the limitation of PPMs. It is difficult to determine whether the effect size of an individual covariate is dependent on its effect on the dependent variable or the interaction between the covariables. However, works on modelling PPMs have little concentration on the multilinearity of variables (Renner et al., 2015).

5.2. Limitations and recommendations

This study takes Geldmann et al. (2016) work as a premise. On the basis of that, we investigated the relationship between observation intensity and associated spatial covariates with more thinkings. They proposed to plot residual from the fitted model against species richness to identify under-/over-sampling areas, which can be further used to improve sampling design. But, residual maps should be interpreted with more caution. The residuals in the model reflect how the model fits. It illustrates the information provided by the model and introduces visible discrepancy which can provide insights into an unusual pattern. And the residuals are calculated by both real observations and dummy points. The predicted pattern does not reflect the natural pattern of the species in the real world, it only depends on the effects

of all spatial covariates. The over-/under-estimation could mean a lack of the function of an unobserved spatial covariate at a particular area (Zuur, Ieno, & Elphick, 2010; Reinhart & Greenhouse, 2018). In our case, the residual map shown in Figure 23 is hard to inspect from a global extent. This is because this study conducted at a large spatial scale while the number of observations is quite limited. The observation points were aggregated to 1 square kilometre grid, and the dummy points were generated at the spatial resolution of 1 square kilometre. This makes it hard to interpret when the entire area is more than 1 million square kilometres. Furthermore, attention should be paid to the explanation of the residual map as well.

PPMs model the probability of each observation to be collected under specific situations. The utility of PPMs brings the advantages not only in finding out the effect levels of each spatial factor, but also helping detecting potential unusual patterns. It focuses on individual observation event and at the same time, each observation is independently influenced by all spatial contexts. Moreover, another advantage that is not discussed in this study brings the potential to explore inter-point interaction among observations. Despite we assumed observations from the partner group are independently collected, it is likely the partner group may assign specific nearby locations to observers which can result in some potential clusters in space. Likewise, volunteers may observe several plants that are not far apart at the same time. On this basis, other types of PPMs can be adopted to analyse inter-point interaction for future work, such as Markov point processes that take both inhomogeneity and interactions into consideration (Berthelsen & Møller, 2008).

Furthermore, the covariates are required to be spatially continuous across the entire space, thus the predicted model needs to be estimated not only on observed points but also on some "artificial" points. This results in the model being estimated at locations without true observations. Such points facilitate the prediction of spatial distribution based on recorded data (Wang & Stone, 2019). Meanwhile, we found some covariables which are seemly insignificant have relatively strong impacts on the predicted pattern. By analysing the dependence of intensity on each covariate, we found covariates with less spatial coverage have larger effects. Therefore, the effect size of each covariate needs to be considered with caution. Also, collinearity among variables is usually ignored before modelling PPMs. This could bring uncertainty to explain the effect size of each covariate. Future work can focus on detecting collinearity by such as principal component analysis (PCA) before estimating the covariates (Zuur et al., 2010).

This study demonstrated the existence of spatial bias in volunteered phenological observations and quantified the influence of potential spatial factors that cause spatial bias. Existing research that explores spatial bias in citizen science projects mainly focuses on species richness (Geldmann et al., 2016; Mair & Ruete, 2016) or environmental monitoring (Millar et al., 2019). The results of this study enrich this finding, indicating that volunteered phenological monitoring is also affected by the uneven distribution of observations.

Ideally, it is recommended to continuously monitor a fixed individual plant at a fixed location. However, it is also possible that some observations have only been observed once. Such observations may be occasionally made by volunteers during their free time. Sparks, Huber, and Tryjanowski (2008) found that around 44% observations are collected at weekends, and volunteers record observations more frequently at weekends. Similarly, Żmihorski et al., (2012) pointed out that there are more observers making observations at weekends. Our study did not consider such temporal bias. Weekday/ weekend bias can also affect the distribution of observations. For example, the arrival of tourists may facilitate the collection of some observations (Tulloch & Szabo, 2012), but this could introduce uncertainty to estimate the effect

of fixed human population density. Future work can separate long-term observations from single observation (only observed once) and also take weekend bias into consideration, to investigate the causes of uneven distribution of observations.

Our results demonstrated road density, human population density, and different land cover affect the distribution of observations in space. But we consider these factors only from the perspective of observer preference and social-economic background. An investigation of volunteer effort in a water monitoring project found that the number of records and monitoring sites are significantly correlated with both population density and education level (Deutsch, 2015). Although our initial exploratory analysis found there are two types of observers – individual and partner group, the effect of two different observers on the number of observations is not included in this study due to the limitations in time and computation. It is possible that some partner groups (e.g., botanical garden) will organise educational activities to make observations in a limited small area (Norfolk Botanical Garden, 2018), which will lead to relatively concentrated observations in space. In addition, since there are some of the partner groups belong to universities or research institutions, different education levels of observers in different regions may also affect data collection.

6. CONCLUSIONS

VGI has been recognized for its timely and low-cost features and has been used in many scientific projects. However, data quality is a complex issue for VGI. In this study, we investigated the spatial bias in volunteered phenological observations from the perspective of volunteers and the local environment. By taking environmental and geographical variables into consideration, spatial bias caused by varying volunteer behavior during the data collection process can be explained.

We estimated the influence level of several spatial covariates to assess the variation in observations. These covariables reflect volunteer activity and preference during the data collection process, which can facilitate future volunteer management. We used seventeen spatial covariates to imply the variation of observation intensity. Our results show that volunteers have a tendency to choose spatial locations for making observations and the data collection process is subject to social-economic factors like human population density and road density. Protocols for effective sampling should take the variation in sampling effort by volunteers into consideration.

Point Process model has been developed as a tool to efficiently assess the presence of observations. It is widely used in modeling species distribution but less explored in phenology studies. PPMs have the potential to investigate the spatial bias in volunteer-collected data. Our results show PPM is a powerful tool to explain the relationship between observation intensity and associated spatial covariates. Furthermore, the exploration of residual gives insights into the unusual spatial pattern of the data collected. But, PPMs also have some limitations. Further research should take the interaction between pair observation into account, as volunteers may choose multiple observation sites nearby. Also, more explanatory variables related to the collection of observations can be considered.

LIST OF REFERENCES

Al-Bakri, M., & Fairbairn, D. (2010). Assessing the accuracy of "crowdsourced" data and its integration with official spatial data sets. In *Accuracy* (pp. 317–320). International Spatial Accuracy Research Association.

Antoniou, V., & Skopeliti, A. (2015). Measures and indicators of vgi quality: An overview. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *2*(3W5), 345–351. https://doi.org/10.5194/isprsannals-II-3-W5-345-2015

Antoniou, Vyron, & Schlieder, C. (2014). Participation Patterns, VGI and Gamification. *Agile 2014*, (JUNE 2014). Retrieved from http://www.geogames-team.org/agile2014/submissions/Antoniou_Schlieder_2014_Participation_Pattern_VGI_and_Gamification.pdf

Arsanjani, J. J., Mooney, P., Zipf, A., Schauss, A., Arsanjani, J. J., Zipf, Á. A., … Mooney, P. (2015). Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets. *Lecture Notes in Geoinformation and Cartography*. https://doi.org/10.1007/978-3-319-14280-7_3

Baddeley, A., Turner, R., Møller, J., & Hazelton, M. (2005). Residual analysis for spatial point processes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(5), 617–666. https://doi.org/10.1111/j.1467-9868.2005.00519.x

Baddeley, Adrian. (2010). Analysing spatial point patterns in R, (Version 4.1), 232.

Basiri, A., & Gardner, Z. (2017). The Impact of Biases in the Crowds ourced Trajectories on the Output of Data Mining Processes.

Beaubien, E. G., & Hamann, A. (2011). Plant phenology networks of citizen scientists: Recommendations from two decades of experience in Canada. *International Journal of Biometeorology*, *55*(6), 833–841. https://doi.org/10.1007/s00484-011-0457-y

Berthelsen, K. K., & Møller, J. (2008). NON-PARAMETRIC BAYESIAN INFERENCE FOR INHOMOGENEOUS MARKOV POINT PROCESSES. *Australian & New Zealand Journal of Statistics*, *50*(3), 257–272. https://doi.org/10.1111/j.1467-842X.2008.00516.x

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Brando, C., & Bucher, B. (2010). Quality in User Generated Spatial Content : A Matter of Specifications Quality of Spatial Content. *13th AGILE International Conference of Geographic Information Science*, 1–8.

Brunsdon, C., & Comber, L. (2012). Assessing the changing flowering date of the common lilac in North America: A random coefficient model approach. *GeoInformatica*, *16*(4), 675–690. https://doi.org/10.1007/s10707-012-0159-6

Camboim, S. P., Meza Bravo, J. V., & Sluter, C. R. (2015). An investigation into the completeness of, and the updates to, OpenStreetMap data in a heterogeneous area in Brazil. *ISPRS International Journal of Geo-Information*, *4*(3), 1366–1388. https://doi.org/10.3390/ijgi4031366

Chandler, M., See, L., Buesching, C. D., Cousins, J. A., Gillies, C., Kays, R. W., … Tiago, P. (2017). Involving Citizen Scientists in Biodiversity Observation. In M. Walters & R. J. Scholes (Eds.), *The GEO Handbook on Biodiversity Observation Networks* (pp. 211–237). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27288-7_9

CitizenScience.gov. (n.d.). About CitizenScience.gov | CitizenScience.gov. Retrieved May 15, 2020, from https://www.citizenscience.gov/about/#

Cleland, E. E., Chuine, I., Menzel, A., Mooney, H. A., & Schwartz, M. D. (2007). Shifting plant phenology in response to global change. *Trends in Ecology and Evolution*, *22*(7), 357–365. https://doi.org/10.1016/j.tree.2007.04.003

Cohn, J. P. (2008). Citizen Science: Can Volunteers Do Real Research? *BioScience*, *58*(3), 192–197. https://doi.org/10.1641/b580303

Delcourt, H., & Delcourt, P. (1996). Eastern deciduous forests. In M. G. Barbour & W. D. Billings (Eds.), *North American terrestrial vegetation* (2nd ed., pp. 357–395). Cambridge University Press.

Dennis, R. L. H., & Thomas, C. D. (2000). Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. *Journal of Insect Conservation*, *4*(2), 73–77. https://doi.org/10.1023/A:1009690919835

Deutsch, W. G. (2015). Trends, challenges, and responses of a 20-year, volunteer water monitoring. *Ecology*

and Society, *20*(3), 14.
Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio- Temporal Point Patterns. Journal of Chemical Information and Modeling* (3rd ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1017/CBO9781107415324.004
Feick, R., & Roche, S. (2013). Understanding the Value of VGI. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (pp. 15–29). https://doi.org/10.1007/978-94-007-4587-2_2
Fritz, S., Fonte, C. C., & See, L. (2017). The role of Citizen Science in Earth Observation. *Remote Sensing*, *9*(4), 1–13. https://doi.org/10.3390/rs9040357
Fry, J. A., Xian, G., Jin, S., Dewitz, J. A., Homer, C. G., Yang, L., … Wickham, J. D. (2011). Completion of the 2006 national land cover database for the conterminous united states. *Photogrammetric Engineering and Remote Sensing*, *77*(9), 858–864.
Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., … Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, *22*(11), 1139–1149. https://doi.org/10.1111/ddi.12477
Gimond, M. (2019, August 9). Point Pattern Analysis. Retrieved May 18, 2020, from https://mgimond.github.io/Spatial/
Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y
Haklay, M. (2013). Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (Vol. 9789400745, pp. 105–122). https://doi.org/10.1007/978-94-007-4587-2
Hefley, T. J., Tyre, A. J., Baasch, D. M., & Blankenship, E. E. (2013). Nondetection sampling bias in marked presence-only data. *Ecology and Evolution*, *3*(16), 5225–5236. https://doi.org/10.1002/ece3.887
Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns. Statistical Analysis and Modelling of Spatial Point Patterns*. https://doi.org/10.1002/9780470725160
Isaac, N. J. B., & Pocock, M. J. O. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society*, *115*(3), 522–531. https://doi.org/10.1111/bij.12532
Jackson, S. P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., & Stefanidis, A. (2013). Assessing completeness and spatial error of features in volunteered geographic information. *ISPRS International Journal of Geo-Information*, *2*(2), 507–530. https://doi.org/10.3390/ijgi2020507
Jacobs, C., & Zipf, A. (2017). Completeness of citizen science biodiversity data from a volunteered geographic information perspective. *Geo-Spatial Information Science*, *20*(1), 3–13. https://doi.org/10.1080/10095020.2017.1288424
Joly, A., Vrochidis, S., Karatzas, K., & Karppinen, A. (2018). *Multimedia Tools and Applications for Environmental & Biodiversity Informatics. Multimedia Tools and Applications for Environmental & Biodiversity Informatics*. https://doi.org/10.1007/978-3-319-76445-0
Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*(2), 401–413. https://doi.org/10.1890/02-5364
Koch, E., Bruns, E., Chmielewski, F., Defila, C., Lipa, W., & Menzel, A. (2009). Guidelines for Plant Phenological Observations, (70), 41.
Koch, E., Bruns, E., Chmielewski, F., Defila, C., Lipa, W., & Menzel, A. (2015). Guidelines for Plant Phenological Observations, (January).
Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139
Mair, L., & Ruete, A. (2016). Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS ONE*, *11*(1), 1–13. https://doi.org/10.1371/journal.pone.0147796
Mateu, J., Usó, J. L., & Montes, F. (1998). The spatial pattern of a forest ecosystem. *Ecological Modelling*, *108*(1–3), 163–174. https://doi.org/10.1016/S0304-3800(98)00027-1
McDonough MacKenzie, C., Murray, G., Primack, R., & Weihrauch, D. (2017). Lessons from citizen science: Assessing volunteer-collected plant phenology data with Mountain Watch. *Biological Conservation*, *208*, 121–126. https://doi.org/10.1016/j.biocon.2016.07.027

Mehdipoor, H., Zurita-Milla, R., Rosemartin, A., Gerst, K. L., & Weltzin, J. F. (2015). Developing a workflow to identify inconsistencies in volunteered geographic information: A phenological case study. *PLoS ONE*, *10*(10), 1–14. https://doi.org/10.1371/journal.pone.0140811

Melaas, E. K., Friedl, M. A., & Richardson, A. D. (2016). Multiscale modeling of spring phenology across Deciduous Forests in the Eastern United States. *Global Change Biology*, *22*(2), 792–805. https://doi.org/10.1111/gcb.13122

Millar, E. E., Hazell, E. C., & Melles, S. J. (2019). The "cottage effect" in citizen science? Spatial bias in aquatic monitoring programs. *International Journal of Geographical Information Science*, *33*(8), 1612–1632. https://doi.org/10.1080/13658816.2018.1423686

Miller-Rushing, A. J., Inouye, D. W., & Primack, R. B. (2008). How well do first flowering dates measure plant responses to climate change? The effects of population size and sampling frequency. *Journal of Ecology*, *96*(6), 1289–1296. https://doi.org/10.1111/j.1365-2745.2008.01436.x

Mocnik, F. B., Mobasheri, A., Griesbaum, L., Eckle, M., Jacobs, C., & Klonner, C. (2018). A grounding-based ontology of data quality measures. *Journal of Spatial Information Science*, *16*(16), 1–25. https://doi.org/10.5311/JOSIS.2018.16.360

Morellato, P., Camargo, M. G. G., Luize, B. G., & Mantovani, A. (2009). The Influence of Sampling Method, Sample Size, and Frequency of Observations on Plant Phenological Patterns and Interpretation in Tropical Forest Trees. In I. L. Hudson & M. R. Keatley (Eds.), *Phenological Research: Methods for Environmental and Climate Change Analysis* (pp. 99–121). https://doi.org/10.1007/978-90-481-3335-2

Nature Today | De Natuurkalender. (n.d.). Retrieved February 26, 2020, from https://www.naturetoday.com/intl/nl/observations/natuurkalender?utm_source=natuurkalender.nl&utm_medium=redirect&utm_campaign=olddomain

Niemi, A., & Fernández, C. (2010). Bayesian Spatial Point Process Modeling of Line Transect Data. https://doi.org/10.1007/s13253-010-0024-8

Norfolk Botanical Garden. (2018). Southeastern Virginia Phenology Network – Norfolk Botanical Garden. Retrieved June 7, 2020, from https://norfolkbotanicalgarden.org/learn/horticulture/southeastern-virginia-phenology-network/

Ostermann, F. O., & Spinsanti, L. (2011). A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management. *Agile 2011*, 1–6.

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*(11), 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Reinhart, A., & Greenhouse, J. (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *67*(5), 1305–1329. https://doi.org/10.1111/rssc.12277

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., … Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366–379. https://doi.org/10.1111/2041-210X.12352

Romo, H., García-Barros, E., & Lobo, J. M. (2006). Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, *29*(6), 873–885. https://doi.org/10.1111/j.2006.0906-7590.04680.x

Rosemartin, A. H., Denny, E. G., Weltzin, J. F., Lee Marsh, R., Wilson, B. E., Mehdipoor, H., … Schwartz, M. D. (2015). Lilac and honeysuckle phenology data 1956-2014. *Scientific Data*, *2*(1), 1–8. https://doi.org/10.1038/sdata.2015.38

Sagl, G., Resch, B., & Blaschke, T. (2015). Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities. *Sensors (Switzerland)*, *15*(7), 17013–17035. https://doi.org/10.3390/s150717013

SEDAC. (n.d.). U.S. Census Grids. Retrieved May 4, 2020, from https://sedac.ciesin.columbia.edu/data/collection/usgrid/methods

See, L., Fritz, S., Dias, E., Hendriks, E., Mijling, B., Snik, F., … Rast, M. (2016). Supporting earth-observation calibration and validation: A new generation of tools for crowdsourcing and citizen science. *IEEE Geoscience and Remote Sensing Magazine*, *4*(3), 38–50. https://doi.org/10.1109/MGRS.2015.2498840

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (Muki). (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, *31*(1), 139–167. https://doi.org/10.1080/13658816.2016.1189556

Sparks, T. H., Huber, K., & Tryjanowski, P. (2008). Something for the weekend? Examining the bias in avian phenological recording. *International Journal of Biometeorology*, *52*(6), 505–510. https://doi.org/10.1007/s00484-008-0146-7

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Tulloch, A. I. T., & Szabo, J. K. (2012). A behavioural ecology approach to understand volunteer surveying for citizen science datasets. *Emu*, *112*(4), 313–325. https://doi.org/10.1071/MU12009

U.S. Census Bureau. (2014). TIGER/Line Shapefiles.

U.S. National Park Service. (2017, December 29). Eastern Deciduous Forest (U.S. National Park Service). Retrieved June 7, 2020, from https://www.nps.gov/im/ncrn/eastern-deciduous-forest.htm

United States Environmental Protection Agency. (n.d.). Ecoregions of North America | Ecosystems Research | US EPA. Retrieved April 30, 2020, from https://www.epa.gov/eco-research/ecoregions-north-america

USA-NPN National Coordinating Office. (2016). *USA National Phenology Network Data Product Development Framework and Data Product Catalog , v 1 . 1*.

USA National Phenology Network. (n.d.-a). Frequently Asked Questions | USA National Phenology Network. Retrieved May 16, 2020, from https://www.usanpn.org/nn/faq#rep_location

USA National Phenology Network. (n.d.-b). Retrieved February 26, 2020, from https://www.usanpn.org/

Van Strien, A. J., Plantenga, W. F., Soldaat, L. L., Van Swaay, C. A. M., & WallisDeVries, M. F. (2008). Bias in phenology assessments based on first appearance data of butterflies. *Oecologia*, *156*(1), 227–235. https://doi.org/10.1007/s00442-008-0959-4

Van Strien, A. J., Van Swaay, C. A. M., & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, *50*(6), 1450–1458. https://doi.org/10.1111/1365-2664.12158

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11), 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Wallace, C., Walker, J., Skirvin, S., Patrick-Birdwell, C., Weltzin, J., & Raichle, H. (2016). Mapping Presence and Predicting Phenological Status of Invasive Buffelgrass in Southern Arizona Using MODIS, Climate and Citizen Science Observation Data. *Remote Sensing*, *8*(7), 524. https://doi.org/10.3390/rs8070524

Wang, Y., & Stone, L. (2019). Understanding the connections between species distribution models for presence-background data. *Theoretical Ecology*, *12*(1), 73–88. https://doi.org/10.1007/s12080-018-0389-9

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Annals of Applied Statistics*, *4*(3), 1383–1402. https://doi.org/10.1214/10-AOAS331

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S. M., … Xian, G. (2018). A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., … Culham, A. (2007). How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, *2*(11), e1124. https://doi.org/10.1371/journal.pone.0001124

Zhang, G. (2019). Enhancing VGI application semantics by accounting for spatial bias. *Big Earth Data*, *3*(3), 255–268. https://doi.org/10.1080/20964471.2019.1645995

Zhang, G., & Zhu, A.-X. (2018). The representativeness and spatial bias of volunteered geographic information: a review. *Annals of GIS*, *24*(3), 151–162. https://doi.org/10.1080/19475683.2018.1501607

Zhu, A.-X., Zhang, G., Wang, W., Xiao, W., Huang, Z.-P., Dunzhu, G.-S., … Yang, S. (2015). A citizen data-based approach to predictive mapping of spatial variation of natural phenomena. *International Journal of Geographical Information Science*, *29*(10), 1864–1886. https://doi.org/10.1080/13658816.2015.1058387

Żmihorski, M., Sparks, T. H., & Tryjanowski, P. (2012). The Weekend Bias in Recording Rare Birds: Mechanisms and Consequencess. *Acta Ornithologica*, *47*(1), 87–94. https://doi.org/10.3161/000164512x653953

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210x.2009.00001.x