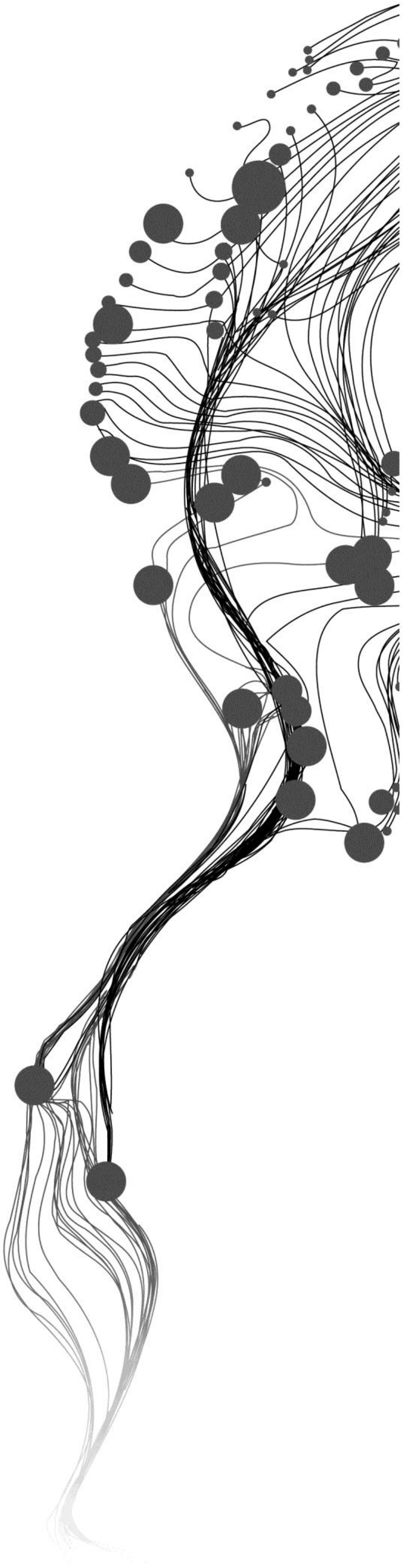
# Identification of Inconsistency and Bias in Volunteered Phenological Observations

HAMED MEHDI POOR
February, 2013

SUPERVISORS:
Dr. R. Zurita-Milla
Ms. Ir. P.W.M. Augustijn



# Identification of Inconsistency and Bias in Volunteered Phenological Observations

HAMED MEHDI POOR
Enschede, The Netherlands, February, 2013

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: Geo-informatics

SUPERVISORS:
Dr. R. Zurita-Milla
Ms. Ir. P.W.M. Augustijn

THESIS ASSESSMENT BOARD:
Prof. Dr. M.J. Kraak (Chair)
Dr. Ir. A.J.H. van Vliet (External Examiner, Wageningen UR, The Netherlands)
Dr. R. Zurita-Milla (Member)
Ms. Ir. P.W.M. Augustijn (Member)

ABSTRACT

Recently, volunteered geographic information (VGI) was introduced as a novel source of data in phenological studies. VGI data can be very useful in phenological research to address environmental phenomena like global climate change. However, quality of this type of data is a major concern to scientists, especially when applying such data for analyses. Presently, quality controlling systems, which check inconsistency and bias in phenological VGI, are primarily based on human capability. Few projects benefit from a computational approach in their checks. In the case of phenological VGI, contextual information like climate data, by posing constraints, can remarkably improve the quality checking process.

The aim of this research is to identify potential inconsistency and bias in phenological VGI by adopting appropriate analysis methods, as well as, proposing a conceptual human/computer workflow for improving phenological VGI. This research proposes a number of methods to identify inconsistency, spatial bias, temporal bias and evaluate the effect of land cover in phenological VGI. The ordered combination of methods are able to: 1) explore the relationship between phenological phases and potential climate variables with exploratory data analysis techniques; 2) define yearly contextual constraints for common species types in The Netherland with regression modelling; 3) filter inconsistent phenological VGI by a constraint satisfaction approach; 4) test complete spatial randomness of the phenological VGI with quadrat counting, as well as, producing empty space distance maps; 5) test the distribution of phenological VGI in the week and land covers, for discovering potential bias, with statistical hypothesis testing.
Further, regarding both implementation results and identified requirements from the user who are interested in phenological VGI quality, a conceptual human/computer workflow is designed. The component and interaction in such the workflow are briefly described, as well.

Eventually, this research revealed that a combination of adopted methods can provide useful information about inconsistency and bias in phenological VGI. However, regarding to the characteristics of species, the effectiveness of methods varies. The results of the methods depend entirely on input parameters of method, which are defined according to the requirements of the user. In fact, this research expands a scope for phenologist to be aware of the quality of phenological VGI. Moreover, for experts in geo-informatics, this research reveals potentials of using a human/computer workflow to control and improve the quality of phenological VGI.

**Key words**: VGI, Phenology, Inconsistency, Spatial bias, Temporal bias, Constraint satisfaction approach and Marked point pattern analysis.

# ACKNOWLEDGEMENTS

Thank you GOD for everything that you have done for me.

First and foremost, I owe my deepest gratitude to my first supervisor Dr. R. Zurita-Milla, for his continuous guidance, motivation and support during this research. I was inspired a lot by his knowledge and constructive comments. You have made me understandable the value of becoming an independent researcher.

My special thanks with high appreciation go to my second supervisor Ms. Ir. P.W.M. Augustijn, for her dedication to always assist me in my thesis and her critical comments aiming to enhance and improve the present thesis. This thesis could not have been written without her elaborate supervision. THANK YOU.

My great gratitude to Dr. Ir. de By who advised me about subject and my acknowledgment goes particularly to Dr. Abkar for his guidance and assistance in the period of course. I would like thank Dr. K. Khoshelham for his support throughout the research period.

All my academic success is based on prayer, encouragement and support of my family. I am very much grateful to my father *Hojjat* and my mother *Tahmineh*. I would like to thank my wife *Sepideh* for her understanding and love. Her support and encouragement was in the end what made this thesis possible.

This acknowledgment cannot be complete without mentioning my friends, Mr. Mahour and Miss Eshraghi, for their friendship, encouragement during the stay in ITC. It was a pleasure to stay and share the joys with you in lovely Enschede.

*Hamed Mehdi Poor,*
*Enschede,*
*The Netherland, February 2013*

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVATION

| | |
|---|---|
| VGI | Volunteered geographic information |
| EDA | Exploratory data analysis |
| CSA | Constraint satisfaction approach |
| CSP | Constraint satisfaction problem |
| SDE | Sum of daily evaporation |
| ICT | Information and communication technologies |
| CSR | Complete Spatial Randomness |

# 1. INTRODUCTION

## 1.1. Phenological monitoring VGI and challenges

Collaborations between scientists and public in data collection is not a new topic (Cohn, 2008). Nevertheless, volunteered geo-information has soared during the last couple of years. This is because information and communication technologies (ICT) and location-aware devices have dramatically changed the way in which volunteers (i.e. non-expert people) can produce geo-information.

As the result, there are several projects around the world that are gathering this volunteered geo-information for monitoring. The most of these volunteer-based projects provide long-term geo-information without considering a specific case study (Dickinson et al., 2010). Table 1.1 lists some of these projects as examples which include domains from investigation to education.

Particularly, the study of periodic animal and plant life cycle events, how seasonal and inter-annual variations in climate affect them and how the abundance and diversity can be modulated by them, is called phenology (Betancourt et al., 2007). The growing interest for understanding the effects of climate change on plant and animal life cycle events leads to the contribution of volunteered geo-information for phenological studies. This is because by volunteered phenological geo-information, both the cost and the time of data collection are significantly reduced. Besides having such advantages, this geo-information allows the investigation of multi-scale research problems (Devictor et al., 2010). For example, the California Phenology Project is a member of a bigger project called National Phenology network in the United States (CPP, 2013), and both address phenological research problems but at a different scale.

In this way, there are several volunteer-based projects, which monitor phenology, in different countries.

Table 1.2 lists some of them. These projects invite volunteers to submit their observations about phenophase (i.e. periodic plant and animal life cycle events) of different category like plants, birds and butterflies.

Table 1.1. Examples of the volunteer-based monitoring projects (Wiggins and Crowston, 2011)

| Type | Name and website | Description |
|---|---|---|
| Investigation | eBird<br>www.eBird.org | Collecting bird observations |
| Conservation | Northeast Phenology Monitoring<br>www.usanpn.org | Monitoring phenology |
| Action | ReClam the Bay<br>www.reclamthebay.org | Restoring local bay's clams and oysters |
| Virtual | Galaxy Zoo<br>www.galaxyzoo.org | Classifying images of galaxie |
| Education | Fossil Finders<br>www.fossilfinders.org | Learning about Devonian fossils |

Table 1.2. Examples of volunteer-based monitoring projects in phenology

| Project name | Description | Web sites |
|---|---|---|
| BudBurst | A network in which people monitor plants as the seasons change through the United States | www.neoninc.org/budburst/ |
| National Phenology Network | A Corporation in the United States that gather different group include citizen scientists, students and government organizations to monitoring the effect of climate change on animal and plants | www.usanpn.org |
| The Natuurkalender | A national educational/scientific observation program which focuses on the identification of the impacts of climate change on annual recurring events in nature. | www.natuurkalender.nl |
| The nature's calendar | A volunteer-based project in which the impact of climate change on United Kingdom wildlife is being monitored regarding seasonal events | www.naturescalendar.org.uk |
| Seasons Observatory | A volunteer-based program to monitor plant bloom and migratory birds | www.obs-saisons.fr |

In practice, the applicability of volunteered geo-information often is subject to quality concerns even though the volunteer-based monitoring projects use varying standards and protocols for the data collection and data analysis phases (Dickinson et al., 2010). This includes error and bias in both the volunteered geo-information itself, and the survey results based on volunteered observations. In this sense, a range of studies has been carried out to check volunteered geo-information quality and quality of corresponding analysis results (Flanagin and Metzger, 2008).

The quality problems associated with volunteered phenological geo-information can be split into two groups, bias and inconsistencies. Bias means any deviation from expected distribution of observations, spatially and temporally. It can be caused by different sources like impact of human population or volunteers' preferences on observations, and they can affect the results from volunteered observations. Besides bias, the inconsistency in volunteered observations is another serious quality issue. Inconsistent observations are not in harmony with other observations, and they can affect the reliability of the analyses and conclusions in phenological studies.

Therefore, a proper understanding of inconsistency and bias in volunteered phenological observations is essential for applying, analysing and deriving conclusions from these data. The following sections describe the problem, objectives, questions, method of this research, as well as, thesis structure.

## 1.2. Research identification

### 1.2.1. Research problem

Early concerns regarding volunteered phenological observations were quality, which includes potential bias and inconsistency in registered observations. They should be identified to ensure usability of these data since they affect inference from volunteered phenological observations.

Currently, the quality checking of volunteered phenological observations is often conducting via human-based quality control systems (i.e. systems using just human capability for checking quality). However, the volume and dimensionality of these data make the quality checking process too overwhelming and time consuming in such systems.

In particular, identifying bias in volunteered phenological observations is a complex since it can spatially and temporally varies among the species, and study areas. For example, since most of the volunteers live

in cities, the concentration of observation locations in city areas can be assumed spatial bias. On the other hand, the absences of observation, in regions where are not species' habitats, cannot be assumed as a bias.
In addition to the bias, climate oscillation and characteristics of volunteered phenological observations can cause inconsistency in observations. For instance, some volunteers may not be able to distinguish species exactly, and as the result, they may interchangeably report those that have similar shapes and sizes. This leads to earlier or later observations than normal. Alternatively, in a local atmospheric zone, where the climate condition differs from the encircling area, some species may do their phenophase earlier or later. Identifying these unusually early or late sightings (compared to other observations in the same region) is another problem that this research addresses.
Besides, there have been no well-designed interactive workflows to identify aforementioned inconsistency and bias by ordering methods and using human, computer and contextual information capabilities. In summary, identifying inconsistency and bias in volunteered phenological observations by using a workflow that relies on human, computer and contextual information in an interactive way is the main problem that this research attempts to address.

### 1.2.2. Research objectives

This research aims at adopting analysis methods to identify potential inconsistency and bias in volunteered phenological observations besides proposing a conceptual human-computer workflow to inform about them in an interactive way. Particularly, the following sub-objectives cover the main objective of this research.

- Find applicable contextual information to support the methods that check for inconsistency in volunteered phenological observations.
- Identify methods to check inconsistency in volunteered phenological observations and to inform concerned users about it.
- Identify methods to check potential bias in volunteered phenological observations and to inform concerned users about it.
- Describe the components and their requirements and interrelationships of a conceptual workflow to inform and improve the quality of volunteered phenological observations.

### 1.2.3. Research questions

According to the sub-objectives, this research attempts to answer following research questions:

- How can contextual information be applied to support methods that check for inconsistency in volunteered phenological observations?
- How can inconsistent volunteered observations be identified in phenological monitoring projects?
- How can potential spatial and temporal bias in volunteered phenological observations be identified and be informed?
- What are the components, requirements and interrelationships in a workflow to inform and improve the quality of volunteered phenological observations in an interactive way?

### 1.2.4. Innovation

The innovation of this research comes from: 1) the use of a novel source of data, volunteered phenological observations, 2) the use of contextual information for identifying potential bias and inconsistency in the data, and 3) the use of human and computer capabilities synergically, to improve the quality of volunteered phenological observations.

## 1.3. Project setup

### 1.3.1. Research Method

To answer the research question posed in sub-section 1.2.3, a research method, which includes the following main steps, is applied:

- **Literature review:** as an initial step in this research, the literatures on the characteristics and quality issues of the volunteered geo-information are reviewed. In addition, this step explores the preceding quality checking efforts, as well as, their advantages and disadvantages. Moreover, the backgrounds of potential methods, which can address research questions, briefly are reviewed in this step.
- **Inspect and prepare available datasets:** analysing available datasets including volunteered phenological observations, climate datasets, land cover datasets and habitat information is taken into account in this step. Additionally, aggregated sub-datasets (which is the result of join of contextual information with volunteered phenological observations) are prepared per species. Moreover, in order to further analysis in next steps, this step produces some new variables.
- **Apply potential methods:** to identify inconsistency and spatial and temporal bias in volunteered phenological observations, the chosen methods, a range from Analysis, modelling and computational methods, are applied. In addition, this step includes the presentation of results from applying methods.
- **Conceptual design of an interactive workflow:** to improve the quality of volunteered phenological observations in monitoring projects, a conceptual interactive workflow is explored. Besides, requirements and interactions of component are described in this step of research.

### 1.3.2. Thesis structure

Totally, this research thesis includes six chapters. Background information, problem statement, research objectives, research questions and research methodology are described as the first introduction chapter. Subsequently, in chapter 2, literature and related works are reviewed from volunteered geo-information to its inherent bias and inconsistency in phenological monitoring. In chapter 3, applied materials and methods are described and are implemented to identify potential inconsistent and biased observations in a real-world datasets. The results and discussions of the method's implementation are presented in chapter 4. Afterwards, an initial conceptual workflow for quality improvement of volunteered phenological observations is proposed in chapter 5. Finally, conclusions and recommendations of this research are presented in chapter 6.

# 2. LITERATURE REVIEW

## 2.1. Introduction

Firstly, this chapter provides an overview of volunteered geo-information, its characteristics and challenges in monitoring projects (section 2.2). Secondly, studies concerned about the quality of this novel data are reviewed to have a background about developed concepts and methods (section 2.3). At the end, an overview is given in the analysis and modelling methods that have been applied or have potential for checking the quality of volunteered geo-information (sections 2.4, 2.5 and 2.6).

## 2.2. Volunteered geographic information

This section explores the use of volunteered geo-information as a new data source as well as its characteristics. The help of volunteers in data collection is not a new notion. In avian studies, it started more than a century ago, with projects like the Christmas Bird Count (2013b). Nevertheless, there has been a tremendous success of volunteered geo-information over the last decades. It is the result of evolutions in ICT like web 2.0 in which users create their own content and sharing via the World Wide Web (Wiersma, 2010).

Furthermore, the development in location-aware technologies like Global positioning system or location based social networks enabled volunteers to produce and share geo-information via ICT (Gouveia and Fonseca, 2008). Indeed, these technologies facilitate registration, storage, presentation, validation and dissemination of what has been called volunteered geographic information (VGI), a term coined by Goodchild (2007).

There are several monitoring projects in the environmental domain that benefit from VGI (Catlin-Groves, 2012). However, they require skilled care and good knowledge of the characteristics of VGI. Features like subjectivity (i.e. based on volunteers' opinion), spatio-temporal attributed, different data capture methods and various accuracy are introduced as the main characteristics that distinguish this data source from the other ones (Deparday, 2010).

For instance, since volunteers record the location and the time of their observations, the VGI is spatio-temporal attributed. Or, there are different methods for enrolling the observations like web, mobile applications and, etc. Further, regarding to the measurement tools the accuracy of observations can vary.

Nonetheless, the most distinguished characteristic of VGI is the subjectivity of the data. VGI is less objective than conventionally acquired data (Tulloch, 2008). In practice, the volunteer is the only person who decides on what, where and when to observe. This is because of factors like age, skills, interests, lodging and experience of the volunteers (Delaney et al., 2008).

Hence, mitigate the effects of subjectivity on the quality of monitoring VGI (i.e. VGI which is used for monitoring purpose) is introduced as a major concern in the environmental studies. Dealing, with this can improve the analysis results in varying research fields using such data (e.g. phenology). The following section explores related works in monitoring VGI quality.

## 2.3. Background efforts into quality checking of Monitoring VGI

Early concerns regarding monitoring VGI were quality issues that potential bias in volunteers' observations is one of them. In this research, term bias means deviation (spatially or temporally) of volunteers from doing a random and homogenous observation through space and time. The bias in monitoring VGI has been addressed in a question-specific manner in different studies because unbiased observations for one research question may not be unbiased for another (Voříšek, 2008). The relevant efforts, which addressed spatial and temporal bias, are described below.

Most of the researchers refer a devotion of observations' locations from a uniform Poisson distribution through the study area as the spatial bias in monitoring VGI. The uniform Poisson distribution, describe the probability of a given number of specific happenings take place in a fixed time interval or space as a discrete probability distribution (Haight and Haight, 1967).

For example, Reddy and Davalos (2003) studied the potential spatial biases due to proximity of monitoring VGI to city, river and road features in Africa. They computed the distance from each of the observations to the nearest feature and then compared the distribution function of distances with generated observations under a uniform Poisson distribution (with an equal number of observations). They did the comparison for each of features separately, and the results showed significant aggregation around the city limits, and along rivers and roads for observation locations.

Ferrer et al. (2006) investigate the birders' (i.e. volunteers watch birds as a recreational activity) spatial preference as the other source of spatial bias in a VGI monitoring project in Catalonia, Spain. By Partial Least-Squares Regression, the relationship between times (days) devoted to bird observations in 10 km × 10 km UTM squares were analysed. In this research, the main pattern showed that large visit frequencies mainly associated with coastal areas. Similarly, in 2010, Boakes et al. Showed how the spatial bias can influence the identification of changes in biodiversity over time as well .

The authors address the uneven distribution of monitoring VGI in a time interval in which volunteers collected data, as temporal bias. Sparks et al. (2008) studied the distribution of volunteers' observations during the week. They examined whether bias exists towards weekend recording in the first spring observations of migrant birds, which called weekend-effect by them. They summarized the numbers of observations in each day for each observation site and then tested by using Analysis of variance. They came to the result that the weekend-effect is mainly a consequence of the greater recorder effort at weekends. However, their results revealed the weekend-effect, to some extent, differed between locations and seasons.

In a similar study, Zmihorski et al. (2012) studied weekend-effect as an inherent temporal bias in biological monitoring VGI. The authors tried to assess potential weekend-effect existed in volunteered observation about the abundance of rare bird species. They used the $\chi 2$ test to determine if the observed number of individuals of each species during weekends and non-weekend days deviated from the expected 2:5 ratio (two weekend days, five non-weekend days). The result of their research also suggests that the weekend-effect reflects only changes in sampling effort across the week in the case of rare bird observations.

Additionally, the preferences of volunteers for making the observation in particular land cover considered as the potential source of bias in monitoring VGI. Harris and Haskell (2007) gave a demonstration of the potential effects of land cover bias on ornithological monitoring VGI. They examined the role of land cover in bird population estimates from roadside observations. The results showed that land cover bias may prevent the detection of bird population changes by roadsides monitoring VGI.

The consistency of observations is considered as another quality issue in monitoring VGI. Initially, the approval of volunteered observations is proposed as a basic strategy for checking the consistency of monitoring VGI in various projects (Bishr and Mantelas, 2008). For example, in the SwiftRiver[1] project volunteers have the possibility to assign values to observations that stand for their trust towards an observation or its observer. The overall computed value for each observation, which shows the consistency of observation, presented as trust value.

[1] http://swift.ushahidi.com

In similar ways, there are some other human-based evaluation processes designed to check and improve the consistency of monitoring VGI. Despite the applicability of these methods, they only consider observation consistency on observation frequency and volunteers who observed them. Hence, unavoidably these approaches fail in first-time and infrequent observations. For example, in cases like disaster response, first-time contributions are rather the rule than the exception and must not be ignored by approval mechanisms.

In another effort in identifying inconsistent observation in monitoring VGI, misidentified observations about birds are filtered out by "eBird" that established filtering method called emergent filtering. This method identifies key periods during a bird's phenology regarding historical VGI and ignore observations are out of the range (Kelling et al., 2011). For different study area, the emergent filtering can be altered due to the difference in regional historical data. However, the use of historically gathered VGI was the only criteria that this method relies on to identify inconsistent observations, and as the conclusion it could not consider the climate-change effect year by year.

Brando et al. (2011) tried to establish comprehensive and adequate specifications (i.e. standards) for checking consistency of monitoring VGI. Nevertheless, inconsistencies, which come from the subjectivity characteristic of VGI, cannot be prevented by defining specifications since their specifications rely on reference data, and it is impossible in many cases. For example, their specifications cannot detect and improve inconsistencies like unusually early or unusually late in phenological monitoring VGI because the observations are bout phenophase and there is no reference observation in each year.

Recently, Schlieder and Yanenko (2010) proposed an approach in which observations confirm each other regarding spatio-temporal proximity as a first criterion and social distance (difference between observers' knowledge level) as a second criterion. Their method introduced a graph in which observations assumed as nodes. For each node, there were edges that connect potential observations to it. They called such a graph as the confirmation graph. For each edge, there was a value (positive or negative) that shows to what extent connected observations accept or deny each other. Then, for each node, a value aggregated to show how that node is consistent. In 2012, they applied a constraint satisfaction approach, instead of aggregation techniques, to address prediction in monitoring VGI in areas that there was incorrect or no observation.

Regarding this literature, following sections explore some methods that have potential for answering research questions posed in this research (see sub-section 1.2.3). They include analysis, modelling and constraint satisfaction approach.

## 2.4. Exploratory data analysis and regression modelling

The contextual information (e.g. climate variables) can play a vital role in improvement of the quality checking of phenological monitoring VGI. Therefore, analysing and discovering their relationship with phenological monitoring VGI is essential. For this purpose, this section explores an analysis method called exploratory data analysis (EDA).

Initially, Tukey defined EDA in 1977, and it is part of  a statistical approach (which mostly employs a variety of graphical techniques) concerned with reviewing, communicating and using less understood data (Kanevski et al., 2009). Today, evolution of statistical computational methods, which rely on computer capability, plays a crucial role in using EDA for researches. In practice, researchers use a mixture of graphical and quantitative techniques of EDA to explore and discover relationships between variables in a problem-solving process.

There are many statistical-computing environments like R featured vastly improved dynamic visualization and modelling capabilities, which allowed scientists to identify outliers, trends and patterns in datasets. In the case of phenology, scientists use EDA techniques extensively to explore the effect of contextual factors on species' life cycles events.

For example, as a first phase in EDA, Gordo et al. (2007) used some univariate and bivariate exploratory graphics (e.g. histogram, scatter plot, and, etc.). They analysed available climate dataset (temperature and rainfall) for studying spring arrival of common swift and barn swallow. In a similar study, De Beurs and Henebry (2004) used EDA method to investigate the effect of temperature on the land surface phenology of Kazakhstan.

In addition, there are the EDA techniques like non-statistical presentations (e.g. geographic maps, space-time cube, and, etc.) that geographic information scientists introduced them as tools for visual exploration of data(MacEachren and Kraak, 2001). These techniques are also useful for qualification of volunteered phenological observations. For example, by using these techniques, Shen (2012) geographically addressed the clustered volunteered phenological observations in The Netherlands.

Most of the time EDA techniques are applied to summarize main characteristics of datasets in an easy to comprehend form with no formulated a hypothesis or statistical modelling usage. However, this research needs a modelling technique since it has a quantitative attitude toward the research problem. This modelling method is the linear regression modelling which briefly reviewed following.

The regression modelling (the term which coined by Francis Galton initially) was used to describe the relationship between the biological phenomenon (Galton, 1890). This statistical technique estimates the relationships among variables. Regression can model dependent variables on independent variable like phenophase time on temperature.

In particular, simple linear regression is one of the earliest cases of regression in which a linear model with a single explanatory variable is estimated by the least squares estimator. The sum of squared residuals of the model is minimized when a straight line is fitted through the set of points, by least squares estimator.

Several phenological studies apply the simple linear regression to model the effect of parameters on species' phenophase. For instance, (Zhang et al., 2007) modelled the responses of vegetation phenophase to rising temperature in the United States. Hudson and Keatley (2009) refer to different phenological studies that use simple linear regression as a tool to model the trends in time series where phenophase dates are plotted against varying time intervals.

## 2.5. Constraint satisfaction approach

In the artificial intelligence field, Constraint Satisfaction Approach (CSA) describes a process that assigns values to variables in a way that certain constraints are satisfied. In this sense, the so-called Constraint Satisfaction Problem (CSP) is defined as a finite set of variables and a set of limitations that confine the domain of variables(Tsang, 1993). For example, the N-queens problem is a well-known CSP among computer scientists. The problem is to place N queens on N different squares in a way that no two queens threaten each other on a chessboard with size of N (Figure 2.1).

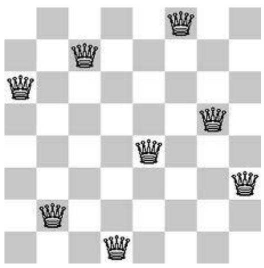


Figure 2.1. A solved N-queens problem

The concept of CSP is used in decision making (Fargier et al., 1995), spatial reasoning (Renz and Nebel, 2007) fields of study to answer scientific research question. Recently, Yanenko and Schlieder (2012) addressed the missing value in geo-tagged reports of sequential events (e.g. phenophases of trees) by simulation of CSP. They used CSA to predict missing events, which previously had been predicted by mean or median estimators in volunteers' observations.

In most of these studies, the visualization of CSP by a graph in which variables are assumed as node and constraints as edges, is identified as helpful sight for problem solving (Russell et al., 1995). The most widely used searching method is "Arc consistency", which tests pairs of variables and checks whether each value of the first variable is consistent with at least one value of the second variable. For solving the CSP in efficient ways, computer scientists have designed and have tested different types of constraint and searching methods. The complete list of them can be found in Tsang (1993), Mackworth (1977) and Bessiere (2006) publications.

## 2.6. Marked point pattern analysis

Most of the phenological monitoring VGI consists of point data, and it is not surprising to find spatial and temporal patterns in such data. In the literatures, the analysis of point data type is introduced as point pattern analysis. In this section, point pattern analysis is reviewed as a method that analyses the potential bias in monitoring VGI.

The term point pattern coined by Hudson and Fowler (1966) as: "the zero-dimensional characteristic of a set of points which describes these points in terms of relative distances of one point to another". In this sense, the phenological monitoring VGI is introducing the point patterns since the locations of observations are registered as x and y coordinates in a reference system.

The analysis of the geometrical structure of patterns formed by points is the main objective of point pattern analysis (Illian et al., 2008). In addition, statistical analyses of the extra information correspond to points, which is called mark, are facilitated by developed analysis method named marked point pattern analysis (Baddeley, 2008).

In marked point pattern analysis, mark (e.g. date of observation) is attached to the point pattern (Rowlingson and Diggle, 1993). Marks can be either a continuous variate (real number) or categorical variate (characters). In this way, temporal attribute also can be analysed and modelled together with the spatial attribute of the point patterns. Further, marked point pattern analysis proposes different tools for spatial and temporal pattern cognition that include visualization techniques, computational methods and statistical hypothesis tests.

# 3. MATERIAL AND METHODS

Chapter 2 reviewed previous efforts related to the checking of inconsistency and bias in monitoring VGI. It also explored some methods that have potential for further development in the quality checking of monitoring VGI. This chapter focuses on the available material and the selected methods in order to answer the research questions.

## 3.1. Data

In this research, two main types of datasets are used. First, a real-world phenological monitoring VGI is selected to be explored for inconsistency and bias. Second, this research benefits from available contextual information, which includes climate, land cover and habitat datasets, to support identification of inconsistency and bias. The source and characteristics of selected datasets are explored in the following sub-sections.

### 3.1.1. Phenological VGI monitoring datasets

The "Natuurkalender" is a Dutch phenological monitoring VGI project that was founded in 2001 as an initiative of Wageningen University and VARA's Early Birds broadcast (Natuurkalender, 2012a). However, other organizations are involved in this program today (Natuurkalender, 2012b).

The main objective of the "Natuurkalender" is to focus on the identification of the impacts of climate change on annual recurring events in nature. For this purpose, anyone who can recognize a few plants or animals is invited to record phenological observations via the website of the project. Up to this time, over 8,000 volunteers and hundreds of school children have participated in this project (Wageningen, 2012).

The observations from the "Natuurkalender" are used in this research as a real-world datasets. The datasets include target species of plants, birds and butterflies. Volunteers submit the date of species' phenophase, in addition to the observation location, which makes such records fall in the domain of VGI. Table 3.1 shows the species and their corresponding phenophase used in this research.

Table 3.1. Target plants, birds and butterflies in this research

| Species name | Phenophase | Category |
|---|---|---|
| **European Pied Flycatcher** | First seen | Bird |
| **Swift** | First seen | Bird |
| **Willow warbler** | First heard | Bird |
| **Common cuckoo** | First heard | Bird |
| **Great tit** | First young fly | Bird |
| **Brimstone butterfly** | First seen | Butterfly |
| **Peacock butterfly** | First seen | Butterfly |
| **Orange tip** | First seen | Butterfly |
| **Wood anemone** | First flower | Plant (herb) |
| **Lesser celandine** | First flower | Plant (herb) |
| **wild chervil** | First flower | Plant (herb) |
| **horse-chestnut** | First flower | Plant (tree) |
| **English oak** | First leaf unfolding | Plant (tree) |

The received datasets contain: 1) unique ID of observations 2) The name of species; 3) city in which observation done; 4) The phenophase name; 5) The date of observation; 6) location of observations which are addressed by x and y coordinates in Dutch National Coordinate System (RD-new); 7) comments that volunteers made about their observations; 8) the year of observation. The geographical extent of the datasets is throughout The Netherlands. The temporal extent of the datasets includes the years 2003 to 2011. Table 3.2 depicts the translated schema of the datasets (originally in Dutch).

Table 3.2. The schema of the "Natuurkalender" dataset

| ID | Species | Phenophase | Date | City | x | y | Comment | Year |
|---|---|---|---|---|---|---|---|---|
| **61** | Brimstone butterfly | First seen | 24-02-03 | Grou | 187000 | 568000 | Male Brimstone in the garden | 2003 |

### 3.1.2. Meteorological datasets

In most phenological studies, meteorological variables are considered as useful contextual information that can help phenologists in their studies. Available meteorological datasets in the geographical extent of the "Natuurkalender" dataset, which is collected by the Royal Netherlands Meteorological Institute (KNMI), are applied in this research.

The KNMI provides weather forecasts and data about climate. In addition, it conducts strategic and applied meteorological research (KNMI, 2012a). Some of this information is available via the data centre part of the KNMI website (KNMI, 2012b). The selected KNMI datasets consist of average daily temperature (ADT) sum of daily precipitation (SDP) and sum of daily evapotranspiration (SDE) in a raster format (1km cell size). More precisely these datasets are interpolation results of daily data collected by about 150 meteorological stations in The Netherlands.

The ADT is the mean temperature of 24 observations in a day, and it is measured in unit degree centigrade. The SDP refers to the daily volume of rain that reaches the ground per square meter, and its unit is kilogram per square meter per second ($kg/m^2$). The SDE means the daily amount of evaporation and plant transpiration from the Earth's land surface to atmosphere (KNMI, 2004).

### 3.1.3. Habitat data

For identifying spatial bias in phenological VGI, it is vital to use the appropriate study area. This is because the lack of observations in regions which are not the natural habitat of a species cannot be assumed as spatial bias. In the "Natuurkalender" datasets, some species (e.g. wood anemone) cannot grow everywhere through The Netherland. For these species, the study area should be restricted to its natural habitat. For this purpose, the Dutch habitat information in "Soortenbank" is applied to confine the study area.

The "Soortenbank" is an initiative of the "ETI" bioinformatics institute and biodiversity collaborating institutions in The Netherlands. Collected habitat information in this project are available to the general public via a website which gives easy access to reliable information on biological diversity (Soortenbank, 2012). In this research, the binary habitat map of wood anemone was extracted and georeferenced to use. The source of the extracted binary habitat map is FLORON foundation, one of the biodiversity collaborating on the "Soortenbank". This data is in the raster format with 25km × 25km resolution (Figure 3.1).

### 3.1.4. Land cover datasets

Most of the species in the "Natuurkalender" datasets have the potential to be seen in the urban environment. The preference of volunteers for doing more observations in urban and built-up areas was introduced as a source of bias in the phenological VGI. This is called the "urban-effect" in this research. Land cover datasets are one of the available datasets that can be applied for studying this bias. In this research, one of the MODIS land cover datasets is applied to test the distribution of observation in the urban area.

The MODIS product is derived from observations ranging one year's input of Terra and Aqua satellite data. In one of five classifications of MODIS products, "MCD12Q1", the 17 land cover classes are defined by the International Geosphere Biosphere programme (NASA, 2012). One of these classes is the urban and built-up area.
The selected MODIS product (Mcd12Q1.005 MODIS combined Aqua and Terra data) has a raster structure with a cell size of 500 meters. The available temporal extent to extract data is from 2003 to 2009 while its temporal granularity is year. In each yearly raster layer, the raster values present the land cover class number. The meaning of each raster value is listed in Table 3.3.

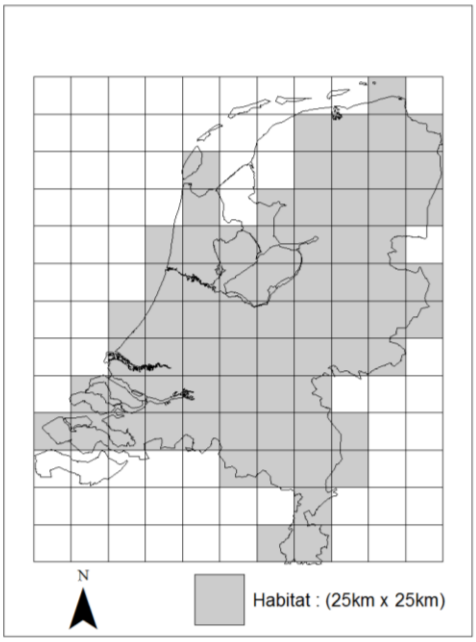



Figure 3.1 Habitat binary map of wood anemone

Table 3.3. The meaning of the raster value in Mcd12Q1.005 MODIS combined Aqua and Terra datasets

| Land cover class | Description |
|---|---|
| 1 | Evergreen needleleaf forest |
| 2 | Evergreen broadleaf forest |
| 3 | Deciduous needleleaf forest |
| 4 | Deciduous broadleaf forest |
| 5 | Mixed forests |
| 6 | Closed shrubland |
| 7 | Open shrublands |
| 8 | Woody savannas |
| 9 | Savannas |
| 10 | Grasslands |
| 11 | Permanent wetlands |
| 12 | Croplands |
| 13 | Urban and built-up |
| 14 | Cropland/natural vegetation mosaic |
| 15 | Snow and ice |
| 16 | Barren or sparsely vegetated |
| 17 | Unclassified |

## 3.2. Methods

The following section gives a description of the methods selected for this research and how they were implemented to address the research questions (see sub-section 1.2.3). Additionally, this section attempts to provide an introduction to the functionality of each of applied methods to explain why they were selected for this research.

Initially, after filtering obvious mistakes from the observations (e.g. observations in which x and y coordinates were recorded 0) the "Natuurkalender" datasets were split per species to sub-datasets. Then, the contextual information are combined with each of sub-datasets. Besides, some additional variables are generated from the available datasets for further analysis.

The second step is to analyse and model the volunteered observations on contextual information in each sub-dataset yearly. That means the highly correlated variable with species' phenophase time will be identified, and their correlation will be modelled. Implementation of these steps enables this research to answer the first research question.

Then, a computational analysis method, which is a constraint satisfaction approach, was applied to identify inconsistent observations. In this step defining and searching of contextual constraints studied in the previous steps is implemented by using computer software and programming language. In this way, the second research question will be answered.

In the third step, marked point pattern analysis identifies potential bias and produce useful information about the spatial distribution of observations, weekend-effect and urban-effect on phenological VGI. The outputs of this step can deliver the interpretable information about the spatial and temporal bias in the observations which enables this research to answer third research question. The order of description of the methods is summarized in Figure 3.2.

.

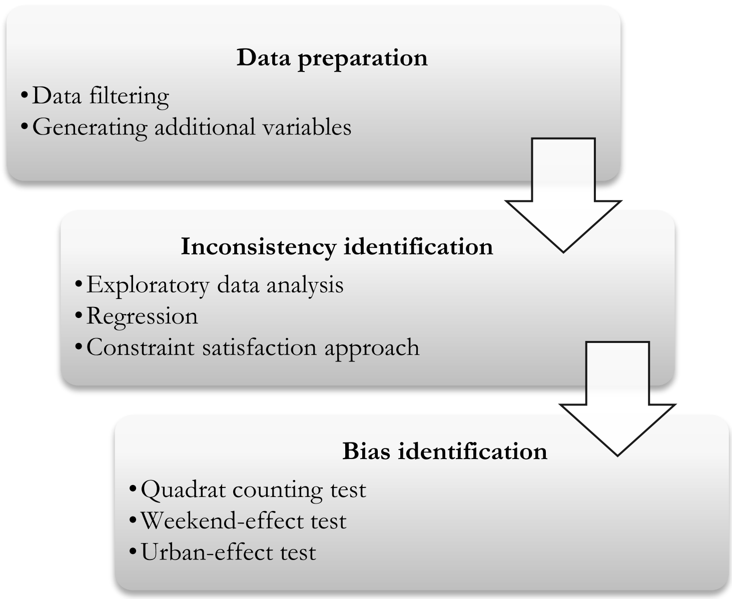



Figure 3.2. The order of selected methods which are to answer research questions

### 3.2.1. Data Preparation

Regarding the characteristics of monitoring VGI, a preparation step is needed to check whether the input dataset is ready for use in the methods. Hence, some preparation tasks have been done to facilitate addressing the research questions. They are summarized in Figure 3.3.

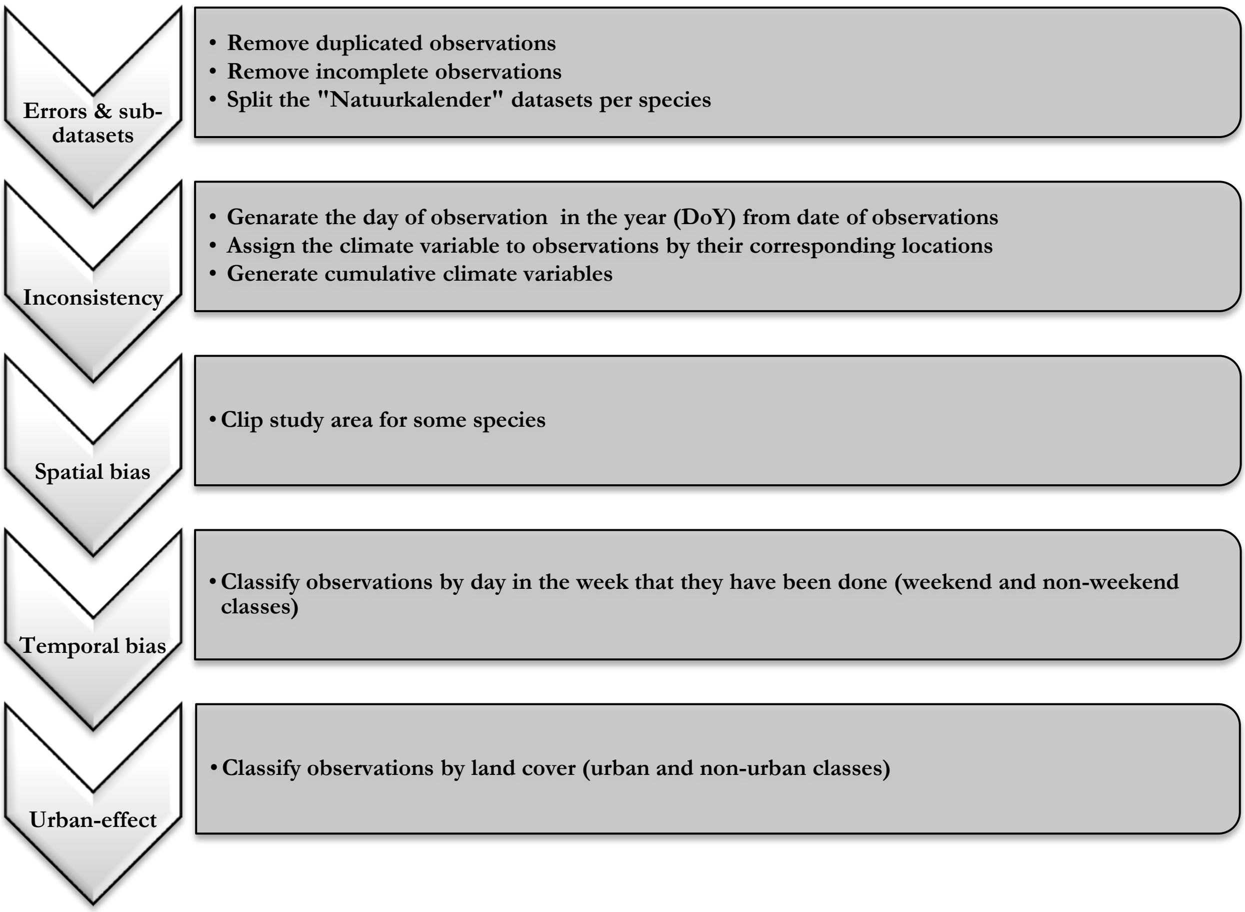



Figure 3.3. The overview of data preparation step

First, duplicated and missed observations (i.e. the locations or dates of observation are missed, or out of feasible/realistic geographical range) were removed from the dataset. Second, the "Natuurkalender" dataset consisting of 2003 to 2009 observations is split into sub-datasets per species.

To meet the third research question about species have natural habitat, habitat binary maps were clipped from the map of The Netherlands, to enclose the study area of this species. This helps to investigate potential spatial bias in the more accurate way.

Phenological studies are interested in timing of phenophases, and most of the species' phenophase were studied in a recurring way (annually). In this sense, the day of the year (DoY) was calculated from the "Date" field for each of observations.

For answering the first and second research questions, urban/non-urban classification of Land cover class (Urban_Nonurabn variable) and weekend/non-weekend classification of Date (Weekend_Nonweekend variable) is generated as new attribute fields to each sub-dataset.

Since the cumulative climate variables are normally as highly correlated factors with the timing of the phenophase (Law et al., 2000, Nizinski and Saugier, 1988, Rutishauser, 2003), these variables are generated from meteorological datasets and are added to the species' sub-datasets by matching observations' locations.

In particular, the cumulative sum of daily precipitation and the cumulative sum of daily evapotranspiration are calculated by summing the amount of SDP and SDE up to DoY of observations (including DoY of observation itself). However, for ADT, phenophase of species are less sensitive to extremely cold days. Hence, a predefined threshold is selected for generating cumulative ADT, above zero degrees centigrade.

### 3.2.2. Inconsistency identification

This sub-section describes which methods are selected to address the inconsistency identification in volunteered observations of the "Natuurkalender" project. In addition, the functionality and implementation of these methods are explained in this sub-section. The ordered overview of used methods is illustrated in Figure 3.4.

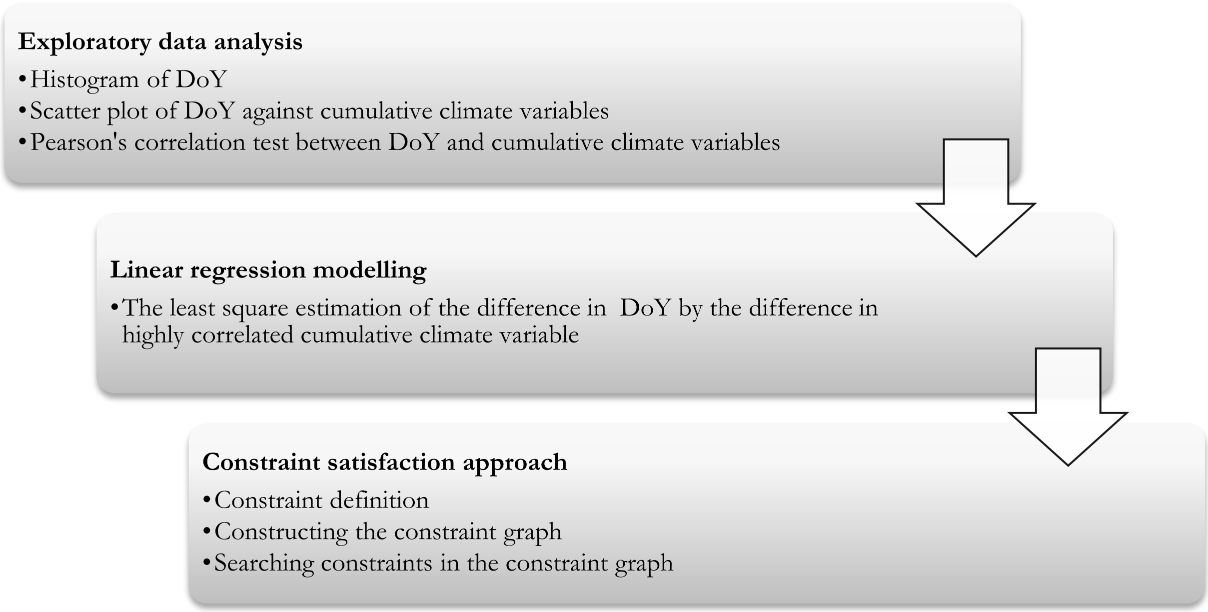



Figure 3.4. The overview of used methods in inconsistency identification step

To answer the first research question, EDA is selected to inspect the annual distribution of the DoY variable and its relationship with climate variables. This method can explore variation in DoY and correlation between variables both graphically and statistically.

For each species, the exploration of the observations started by plotting the histogram of the DoY of the phenophases for in each year. It helps to see an overview of the distribution of observations through the year. Scatter plots are provided for exploring the qualification of correlation between the DoY and the cumulative climate variables. In addition, a correlation test is conducted to quantify the correlations. In this research, Pearson's correlation coefficient between DoY and each of cumulative climate variables are calculated. The covariance of the two variables, divided by the product of their standard deviations, is the exact meaning of this coefficient.

The above-mentioned EDA methods are applied on yearly observations and yearly classified observations according to land cover classes (urban/non-urban). This helps to understand the relationship between observations in each case, in addition to, provide evidence for choosing the supportive variable for defining constraint.

Regarding the result of EDA methods, the differences in DoY, which were calculated between all pairs of observations, was modelled on their difference in highly correlated variable in each year. The simple linear regression was used for this purpose. This method uses least squared estimator to find coefficients of a line best fitted to the point of scatter plots of the differences in DoY against the differences in highly correlated variable. In fact, the estimator minimizes the sum of the vertical distances from the line to points of the scatter plot (Figure 3.5).

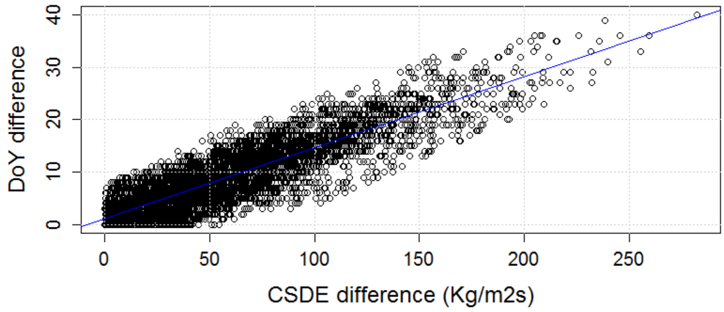



Figure 3.5. An example of the best fitted line by least square estimator

The EDA and simple linear regression methods are implemented in the R statistical software and the complete results of plants, birds and butterflies can be found in the Appendix A. Consequently, the outcomes of these methods are used in further steps of the inconsistency identification process, CSA.

For checking inconsistency in the "Natuurkalender" observations, the CSA (see section 2.5) is used. In this research, the method searches the satisfaction of a constraint through the observations are linked to each other and identifies linked observations that cannot satisfy the constraint. The main steps of this method are described in the following.

The slope of the fitted line (from regression modelling) is used to define a constraint in each year per species. In practice, the constraint is a range function that determines how the two observations can confirm the timing of their phenophases according to their difference in the highly correlated cumulative climate variable. As an example, Figure 3.6 shows how the three observations confirm or disconfirm each other when the correlated climate variable is cumulative SDE.

In this example, $O_1$, $O_2$ and $O_3$ are three volunteered phenological observations. The $\beta_1$ is the extracted slope in the year observations have been done, and $\beta_0$ is maximum variation that under exactly the same condition of the cumulative SDE may happen in DoY of the species. The $\beta_0$ can be introduced by experts (e.g. biologists). The $\Delta$ shows the difference in cumulative SDE.

Take $O_1$ as the observation that wants to participate in the constraint checking with $O_2$ and $O_3$. First, the potential difference in DoY should be predicted by the extracted linear model regarding the difference in highly correlated climate variable ($\beta_1 \times \Delta$). Then, two limited ranges are defined according to the predicted difference in DoY and $\beta_0$. If the real difference in DoY exceeds the ranges, the pair is introduced as consistent ($O_1$, $O_2$). On the other hand, the pair $O_1$ and $O_3$ introduces inconsistency since their difference exceeds the range.

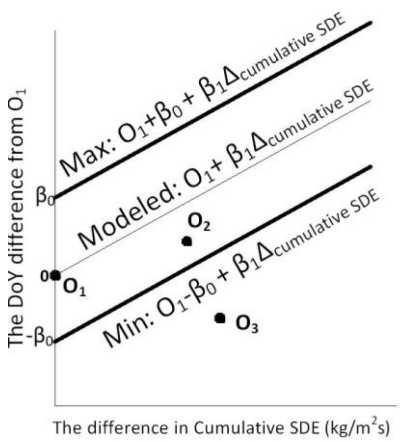



Figure 3.6. A defined contextual constraint based on the differences in cumulative SDE.

As a rule of thumb, in spatial proximity through The Netherlands, the climate variables like SDT can be assumed almost the same. In this research, the spatial proximity was used to link observations as a graph in which observations are assumed as the nodes. For defining proximate observations in the "Natuurkalender" datasets, the x and y coordinates of the observations are used to construct a graph in which observations nearer than a specified threshold (e.g. 40km) are connected to each other. This graph can reduce the impact of other environment variables since the defined constraint just relies on cumulative climate variables. The Delaunay triangulations where all edges longer than a threshold are removed used to create such a graph for each year per species.

After searching observation pairs, the observations identified as inconsistent were checked based on the comments provided by volunteers. This was done by selecting from the datasets a number of observations that based on comments, are potentially inconsistent. The check was performed to identify how many of these sightings were indeed found by applied method in this sub-section.

For both graph construction and constraint checking through the graph, a customized tool was designed by model builder of ArcGIS 10.0. This tool was implemented on each sub-datasets (annually) to create graph information (i.e. observations and connections), and the constraint checked for each pair by Python programming language. The inconsistent pairs were selected and visualized manually in ArcGIS (see sub-section 4.3.3).

### 3.2.3. Bias identification

In this sub-section, methods applied on the "Natuurkalender" data are described to show how they can identify spatial and temporal bias, as well as, urban effect. Firstly, to check the interaction between the observations (e.g. denser observations around cities than in the countryside) through the study area the Complete Spatial Randomness (CSR) of them was tested. By considering the potential study area, this test was conducted on yearly observations. The hypothesis of CSR for a point pattern declares that: the number of observations in the study area follows a uniform Poisson distribution with given mean count per uniform subdivision (which is known as intensity or $\lambda$).

One of the classical tests of CSR is quadrat counting test which is based on statistical hypothesis testing. In this test, after dividing study area to subregions "quadrats", the distribution of the number of observations fallen in quadrats were compared with counting under the CSR condition. For implementing this test, the study area was divided into equal size quadrats. The number of observations that fall in each quadrat is counted, and the Pearson's $\chi^2$ test is used to check the validity of the null hypothesis.

The Pearson's $\chi^2$ test is relied on a null hypothesis which states that the frequency of a particular phenomenon observed in a sample (e.g. number of observations done in each quadrat) is consistent with Pearson's $\chi^2$ distribution. This is assumed that these events cannot happen at the same time, and the sum of their probabilities should be 1.

For example, Figure 3.7 illustrates a point pattern with the rectangular study area. In this case, 78 observations are divided among 6 squares. The "theoretical frequency" for any cell under the null hypothesis of the CSR was calculated by the equation (1), and the statistic of the test was calculated from the equation (2):

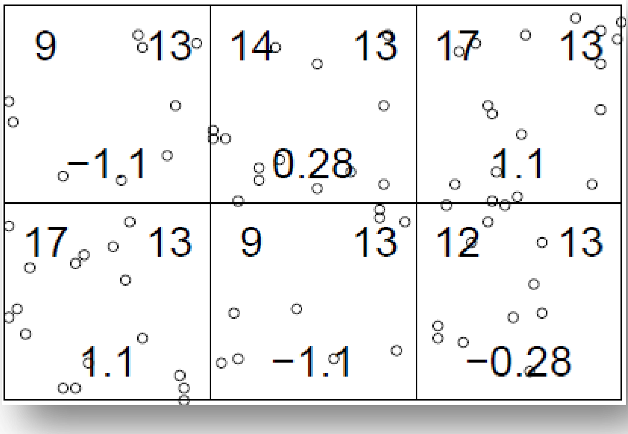



Figure 3.7. The quadrat counting of an example point pattern (Baddeley, 2008)

$$\lambda_i = \frac{Number\ of\ observations}{number\ of\ cells} \tag{1}$$

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - \lambda_i)^2}{\lambda_i} \tag{2}$$

Where:
$X^2$: Pearson's cumulative test statistic, which asymptotically approaches a $\chi^2$ distribution;
$\lambda_i$: An expected (theoretical) frequency, asserted by the null hypothesis (mentioned at the top right of each cell in Figure 3.7);
$O_i$: An observed frequency (mentioned at the top left of each cell in Figure 3.7);
n: The number of cells.

The p-value (i.e. the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true) can be calculated by comparing the value of the statistic to a $\chi 2$ distribution. In this example, the calculated test static is 5.07, and as a result the p-value, which is equal to 0.40, can be extracted (under the chosen 95% confidence interval). Finally, the Pearson's residuals, which used for detecting the quadrats that have the most effect on the result of test (mentioned at the bottom of each cell), are calculated as:

$$Pearson\ residual = \frac{O_i - \lambda_i}{\sqrt{\lambda_i}} \tag{3}$$

The test of CSR is conducted for observations in each year per species to quantify the deviation from a uniform Poisson distribution. However, the main critique to the result of the quadrat counting test is the lack of interpretable information about spatial bias. This test has a rigid attitude by saying yes or no to the question "if a uniform Poisson distribution exists in the observations". Hence, tangible information is needed to qualify the spatial bias in the observations.
To compensate this defect, a complementary method is conducted to qualify the spatial bias in a meaningful way for all users. This is the mapping of empty space distances. The distance from fixed reference locations on a fine grid in the study area to the nearest observation in a point pattern is

calculated. This distance is called the empty space distance (Baddeley, 2008). Mapping these empty space distances based on a user required grid size can provide visually interpretable information for all users who are interested in phenological monitoring VGI. In this research, empty space distances are mapped by using 1km x 1km reference grids (an assumption).

Secondly, the weekend-effect bias is introduced as a common temporal bias, which can affect phenological monitoring VGI data quality (see section 2.3). For quantifying the weekend-effect, this research selected Pearson's $\chi^2$ test of the distribution of the day of observations in the weekend and non-weekend days. It tests the hypothesis "Weekend_Nonweekend" variable is fair (i.e. all seven days have an equal chance of doing observation by volunteers in them). Since the domain of the variable is weekend and non-weekend, the probability of domain values assumed 2/7 and 5/7, respectively. This is because there are 5 ordinary days and 2 weekend days in a week. Thirdly, in a similar way, this test is conducted for "Urban_Nonurban" variable to investigate the effect of urban land cover on the "Natuurkalender" data. However, the domain of this variable differs from species to species.

All methods described in this sub-section are implemented in the R statistical software. In particular, the "spatstat" module was used to convert the prepared sub-datasets per species to a marked spatial point pattern (an object of class "ppp": planner point pattern). In addition, the boundary of the study area of each species was imported as a polygon, which includes polygonal holes, by "maptools" module (an object of class "Owen": the study areas that are presented as a window in two-dimensional space). The corresponding R-scripts are presented in Appendix B.

# 4. RESULTS AND DISCUSSIONS

## 4.1. Introduction

A number of experiments are conducted on the Dutch phenological VGI to test the proposed methods. The phenological VGI dataset contains a variety of species, and before performing the experiments the characteristics of the species were evaluated in order to select the best sub-set of representative species (section 4.2) for the experiments that were performed.

In general, the inconsistency experiment consists of three steps: selection of the climate variable, finding inconsistencies for two selected species and validation of the results with available information. In sub-section 4.3.1 and sub-section 4.3.2, the result of exploring distribution of phenophase time and its correlations with cumulative climate variables are presented for the chosen species. The results of applying the CSA to identify inconsistency are presented, compared and discussed in sub-section 4.3.3. In addition, the observations identified as inconsistent were reviewed regarding the comments provided by volunteers in this sub-section. Sub-section 4.3.4 presents the inputted parameters to the CSA, as well as, sensitivity of the method to these parameters.

Finally, section 4.4 discusses the results from the implementation of methods that identify spatial bias, temporal biases and urban-effect. In addition, the results of empty space distances production are shown to illustrate how such raster layer can inform and help users about preventing spatial bias.

## 4.2. Selection of species

In this research, three distinctions are crucial about selection of species: 1) is the species expected to be highly correlated to climate variables? 2) can the species be observed in the daily environment of the volunteer? and 3) does the species have a fixed habitat? All species included in the “Natuurkalender” are common species, but there are clear differences that will influence the selection of species for presenting the result of inconsistency identification and bias identification. The selected species are summarised in Table 4.1, and the reasons for selection are illustrated in the following of this section.

Table 4.1. Selected species for presenting the results

| | Inconsistency identification | Bias identification | | |
|---|---|---|---|---|
| | | Quadrat counting test | Urban-effect test | Weekend-effect test |
| Wood anemone | × | × | × | × |
| Swift | × | | × | × |
| Great tit | | | | × |

When finding inconsistencies the method uses climate variables. Selection of the species to be used, sensitivity to climate variables is a vital aspect to consider. In general, plants are expected to have the most correlation to climate variables and birds are expected to be less sensitive. To present the sensitivity analysis, wood anemone and swift are selected as examples.

The spatial bias identification can be conducted for all species; however, for species that have specific habitat; this should be investigated within the study area that is restricted to the natural habitat of the species. The wood anemone cannot be found in some part of The Netherlands; therefore, it is selected as the generic case study in spatial bias identification.

Temporal bias can only be expected (or expected in a more pronounced way) for species that are not living close to the observers. For these species, the volunteer will have to go on a walk, or biking trip and these trips are undertaken more frequently during weekends. For detecting temporal bias, three species were selected. The great tit and swift, common birds that presents in urban areas and for which no temporal bias is expected and wood anemone, a plant with a non-urban habitat and, hence, more likely to be observed during weekend trips.

For all species in the "Natuurkalender", there is no excuse for volunteers to do their observations more in urban land than other possible land covers. In this sense, all species can be selected for urban-effect identification. The wood anemone and swift are selected, as the example, to present the result of urban-effect test in this research. Both are common species, and the urban area is not their natural habitat.

## 4.3. Inconsistency identification

### 4.3.1. Selection of the environmental variable

In order to perform inconsistency checks, the first step is to select an climate variable. In this research, three potentially suitable climate variables are checked: cumulative ADT, SDP and SDE. Correlation between these variables and phenophase time of selected species were tested to find the most correlated variable. The phenophase time and its relationships with cumulative climate variables are explored by using three different methods: histogram, scatter plots and the Pearson's correlation test.

Figure 4.1 displays the histograms of the DoY variable for wood anemone and swift for 2007. The variability in DoY for swift is less than for wood anemone. This comparison in all years, 2003 to 2009, shows the same results. This result can be related to the mobility of the swift that compresses this phenophase (first seen) to a shorter time interval. The wood anemone is fixed and have natural habitat, and its phenophase (first flowering) takes place in a longer time interval. This is because they are more sensitive to environment variables, which can vary over different parts of the study area. Alternatively, this variation can be resulted from weekend-effect bias which explained in 4.4.2.

For visually exploring the correlation between the DoY variable and the cumulative climate variables, the scatter plots of these species for 2007 are plotted as an example (Figure 4.2 and Figure 4.3). The scatter plots in this year, and also in other years showed that both cumulative ADT and cumulative SDE are highly correlated variables with the phenophase time of these two species. However, more precise examination has been done by the Pearson's correlation test and the results are listed in Table 4.2 and Table 4.3. The correlation coefficients prove that both cumulative ADT and cumulative SDE are highly correlated variables with the phenophase of this species in all years.

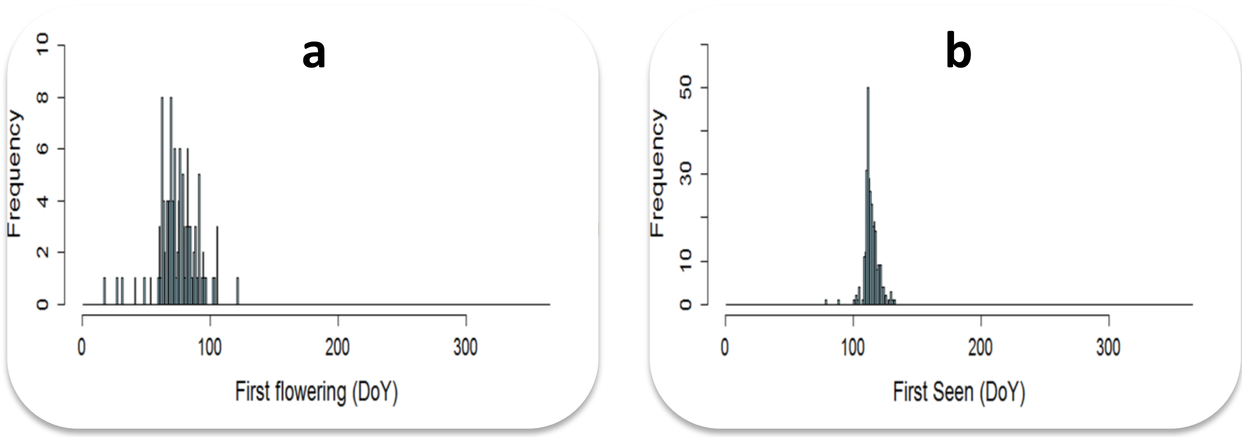



Figure 4.1. The histogram of DoY variable in 2007 a. wood anemone, b. swift

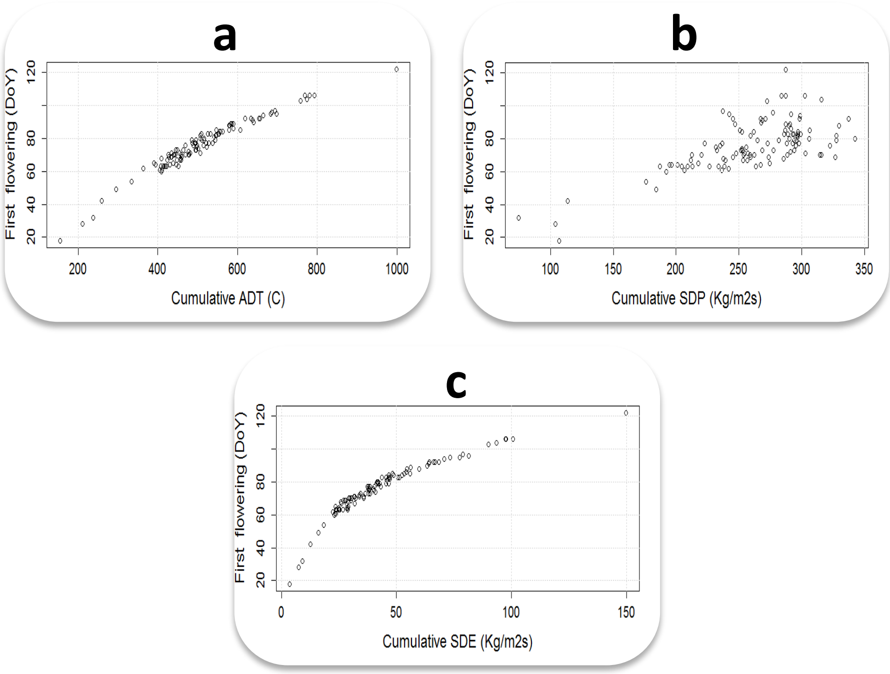



Figure 4.2. The time of first flowering of wood anemone against a. cumulative ADT, b. cumulative SDP and c. cumulative SDE (2007)

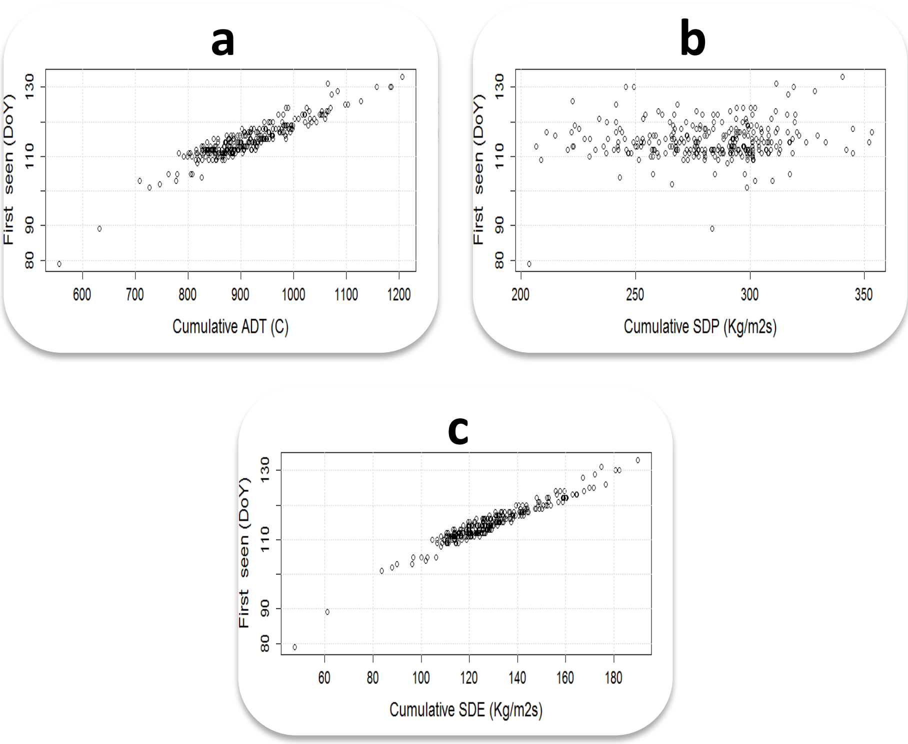



Figure 4.3. The time of first seen of swift against a. cumulative ADT, b. cumulative SDP and c. cumulative SDE in 2007

Table 4.2. Pearson's correlation coefficients between first flowering of wood anemone and cumulative climate variables

| | | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|
| **Yearly observations** | Cumulative ADT | 0.94 | 0.91 | 0.96 | 0.98 | 0.97 | 0.95 | 0.95 |
| | Cumulative SDP | 0.53 | 0.70. | 0.96 | 0.93 | 0.92 | 0.93 | 0.73 |
| | Cumulative SDE | 0.96 | 94 | 0.97 | 0.99 | 0.97 | 0.95 | 0.94 |

Table 4.3. Pearson's correlation coefficients between first seen of swift and cumulative climate variable

| | | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|
| **Yearly observations** | Cumulative ADT | 0.79 | 0.85 | 0.91 | 0.84 | 0.91 | 0.88 | 0.97 |
| | Cumulative SDP | 0.57 | 0.18 | 0.56 | 0.15 | 0.08 | 0.36 | 0.53 |
| | Cumulative SDE | 0.86 | 0.94 | 0.99 | 0.86 | 0.95 | 0.94 | 0.97 |

For both species, for most of the years the correlation coefficient related to cumulative SDE is a little higher than the cumulative ADT. Moreover, for wood anemone this variable shows stronger correlation than for swift (as was expected). For swift, the averages of correlations coefficients, with cumulative ADT, in these years are 0.86 while this is 0.91 for wood anemone. In similar calculation, for swift, the average of correlation coefficients with cumulative SDE, from 2003 to 2009, is 0.92, but for wood anemone is 0.97 ,which is a higher correlation.

#### 4.3.2. Urban versus Non-Urban

Inconsistency checks can be performed on the complete dataset, however, for selected species; urban observations are likely to deviate from non-urban observations. This could be due to sensitivity to the different micro climate in cities. Local temperature can deviate from a few degrees centigrade leading to a different phenophase time compared to surrounding non-urban area (more exposed areas). In this case, it would make sense to split the dataset into two subsets before performing the inconsistency checks. A simple way of checking this is to conduct the Pearson's correlation test for the urban and non-urban land cover separately, and check if these subsets lead to different results. The results are listed in the Table 4.4 and Table 4.5.

Table 4.4. Pearson's correlation coefficients between first flowering of wood anemone and cumulative climate variables, according to observations' land cover

| | | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|
| **Yearly observations in the urban land cover** | Cumulative ADT | 0.89 | 0.91 | 0.92 | 0.99 | 0.96 | 0.97 | 0.95 |
| | Cumulative SDP | 0.59 | 0.7 | 0.4 | 0.97 | 0.67 | 0.79 | 0.65 |
| | Cumulative SDE | 0.97 | 0.95 | 0.94 | 0.99 | 0.92 | 0.98 | 0.94 |
| **Yearly observations in non-urban land cover** | Cumulative ADT | 0.95 | 0.92 | 0.96 | 0.92 | 0.98 | 0.96 | 0.95 |
| | Cumulative SDP | 0.52 | 0.7 | 0.97 | 0.55 | 0.94 | 0.93 | 0.76 |
| | Cumulative SDE | 0.96 | 0.94 | 0.97 | 0.94 | 0.97 | 0.96 | 0.93 |

Table 4.5. Pearson's correlation coefficient between first seen of swift and cumulative climate variable, according to observations' land cover

| | | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|
| **Observations in the urban land cover** | Cumulative ADT | 0.79 | 0.89 | 0.92 | 0.83 | 0.91 | 0.86 | 0.97 |
| | Cumulative SDP | 0.58 | 0.1 | 0.52 | 0.21 | 0.19 | 0.38 | 0.57 |
| | Cumulative SDE | 0.87 | 0.94 | 0.99 | 0.87 | 0.95 | 0.93 | 0.98 |
| **Observations in non-urban land cover** | Cumulative ADT | 0.8 | 0.86 | 0.91 | 0.85 | 0.92 | 0.9 | 0.97 |
| | Cumulative SDP | 0.57 | 0.23 | 0.62 | 0.09 | 0.02 | 0.35 | 0.51 |
| | Cumulative SDE | 0.86 | 0.94 | 0.99 | 0.86 | 0.96 | 0.95 | 0.97 |

The results of Pearson's correlation test in both urban and non-urban land covers show that cumulative ADT and cumulative SDE are highly correlated variables. There is no significant difference in correlation coefficients calculated from classified observation and all observations together. Then, for further steps, the cumulative SDE is selected for using in the CSA. All annual observations are used to model the potential correlation because of small differences in the correlation coefficient with observations done only in the urban and non-urban areas.

Next, by taking all possible pairs in annual observations, the differences in phenophase time (DoY) are calculated in between as well as the difference in cumulative SDE. Figure 4.4 shows scatter plot of the differences in DoY, calculated between every possible pair of wood anemone and swift observations, against their difference in cumulative SDE in 2007.

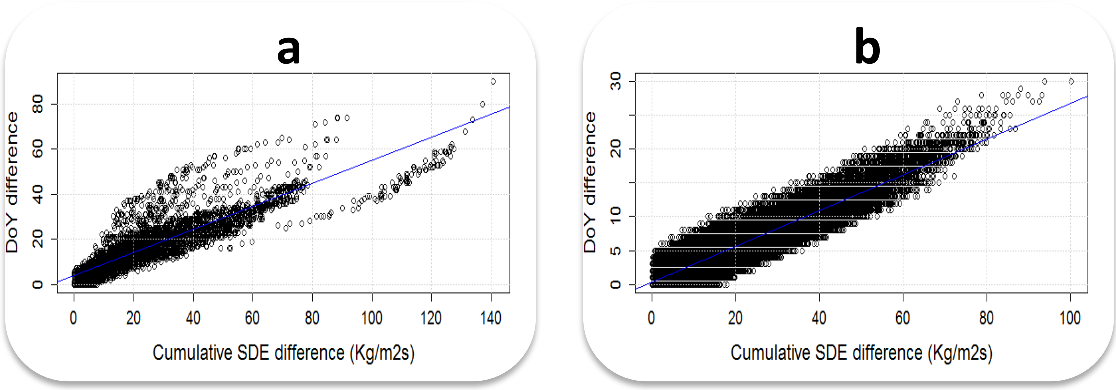



Figure 4.4. The differences in DoY (for every possible pair of observations) against their corresponding differences in cumulative SDE in 2007 a. wood anemone ans b. swift

Then, modelling is done by the least square estimator. In fact, the relationship between the difference in phenophase time with respect to the difference in cumulative SDE is extracted for both wood anemone and swift (Table 4.6). As it was expected, the plants are more sensitive to the cumulative variable because the slope ($\beta_1$) for wood anemone is larger than swift in all years.

Table 4.6. The slope of fitted line to the difference in phenophase time against the difference in cumulative SDE

| | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|
| Wood anemone | 0.61 | 0.80 | 0.55 | 0.48 | 0.51 | 0.86 | 0.59 |
| swift | 0.27 | 0.32 | 0.42 | 0.26 | 0.26 | 0.28 | 0.36 |

### 4.3.3. Identified inconsistent pairs of observations

This sub-section shows how the CSA deals with inconsistency identification in the "Natuurkalender" observations. For this purpose, two species were selected; wood anemone, a fixed species that is highly dependent on climate variables, and swift, a moving species is less dependent on climate variable. For each of these species, inconsistencies were identified and compared to investigate how the method works for them.

To assess the "correctness" of the identified inconsistencies, reference data (i.e. reliable information that show the exact time of phenophase through the study area) would be needed, but this is not available in the case of "Natuurkalender" observations. Nonetheless, the comment registered by volunteers could help to explore if the proposed method is efficient in this research. Volunteers entered comments like "Very many weeks earlier than usual!", "may be this is the last blooming in this year", "first flower, never seen so early". Such comments were used to evaluate if the method could identify them as inconsistent observation.

For consistency checking of wood anemone, a maximum of one week variation is assumed, which means that plants that have similar evaporation conditions cannot vary more than seven days in first flowering. For swift, this variation is set to four days. This is because the phenophase time interval for swift is shorter (see sub-section 4.3.1).

Regarding the result of linear regression, the constraint propagation was implemented on yearly observations. The inconsistent pairs of observations were identified in a 40km proximity, which means that observation pairs located within this distance participated in the constraint checking process. The percentages and number of inconsistent observations for wood anemone and swift in each year are listed in Table 4.7.

The number of inconsistent pairs is low in all years, varying from 0 for wood anemone in 2005, to 17 for swift in 2008. There seems to be no relationship between the percentages of inconsistency between the two species. Wood anemone shows more variation compared to swift, but this can be explained because it is more correlated to climate variables and weather patterns differ from year to year. The weather in spring 2005 was very cold. KNMI reported the night of March 3$^{th}$ to 4$^{th}$ a temperature of -20 degrees centigrade which was the coldest March ever. In our experiments no inconsistent observations were found in this year.Spring in 2007 was extremely warm and sunny; this can perhaps explain the relatively high percentage of inconsistency (11%) for wood anemone. Perhaps this has led to early local flowering of wood anemone, leading to very early observations.

As an example, Figure 4.5 and Figure 4.6 depict identified inconsistent pairs for selected species in 2008. The day of observations in the year (DoY) is plotted next to the observations in both figures, as well. The standalone points for which two numbers were assigned, are informing about two observations done in the same city (or postal code) but at two different dates. Those are introduced as inconsistent observation by the method.

Table 4.7. The percentage and number of observations which were identified as inconsistent by CSA

| | | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|---|
| wood anemon | Number of inconsistent observations | 3 | 5 | 0 | 4 | 13 | 10 | 5 |
| | Percentage of inconsistent observations | 4% | 6% | 0 | 3% | 11% | 8% | 4% |
| | Total number of observations | 67 | 83 | 105 | 123 | 117 | 118 | 109 |
| swift | Number of inconsistent observations | 8 | 4 | 2 | 14 | 4 | 17 | 8 |
| | Percentage of inconsistent observations | 6% | 1% | 1% | 6% | 1% | 7% | 3% |
| | Total number of observations | 123 | 245 | 265 | 227 | 302 | 239 | 202 |

Each identified pair includes inconsistency which does not mean that one of its observations is surely mistaken. For example, in the northeast part of The Netherlands, a volunteer reported first flowering of wood anemone in 23$^{th}$ of January in the city Emmen, which undoubtedly is inconsistent (Figure 4.7). It can be either a mistake or a blooming that probably followed by a colder period in which the flowers disappeared, and they re-emerged in March.


Figure 4.5. The pairs which were introduced as inconsistency for wood anemone (2008), and numbers plotted next to each observation shows the day of observations in the year (DoY)

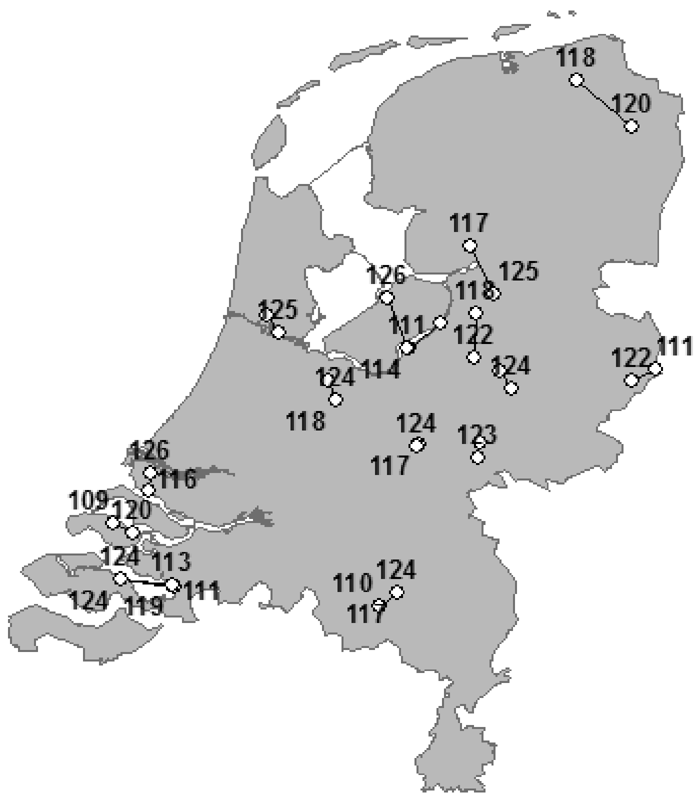



Figure 4.6. The pairs which were introduced as inconsistency for swift (2008), and numbers plotted next to each observation shows the day of observations in the year (DoY)

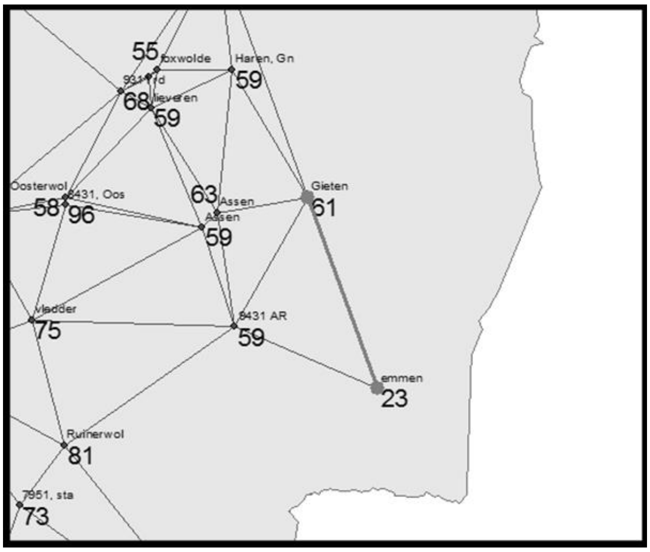



Figure 4.7. The time of first flowering of wood anemone reported, in 23th of January, in the city of Emmen, in 2007, which is identified as an inconsistent observation by the CSA

For wood anemone in 2008, all observations with comments indicating inconsistency were extracted and checked to know whether they were identified as inconsistent observations by the CSA. Table 4.8 illustrates which of them have been identified as the inconsistent observations by "+". Although there were many comments for swift, none of them could clarify which observations could be mistaken or inconsistent.

Table 4.8. All observations that founded inconsistent regarding volunteers' comments in 2008

| ID | Species | Date | City | X | Y | Comment | |
|---|---|---|---|---|---|---|---|
| **369** | Wood anemone | 26-01-08 | Enschede | 257000 | 472000 | Very many weeks earlier than usual! | |
| **885** | Wood anemone | 10-02-08 | Wageningen | 174100 | 444960 | First flower. Never seen so early. | + |
| **1048** | Wood anemone | 17-02-08 | Roosendaal | 88000 | 392000 | very early On the same spot in the garden as previous years | + |
| **4387** | Wood anemone | 26-10-08 | Gemeente | 215000 | 457000 | Probably last flowering. In shoulder along footpath on Sunday morning at 9.00 was observed. In spring bloom much more wood anemone. | + |

#### 4.3.4. Sensitivity to input parameters

The method as introduced (see sub-section 3.2.2) has a number of input parameters that can influence the results. One of these input parameters is the maximum feasible variation in phenophase time under the same cumulative SDE. For wood anemone, this was set to a week as an assumption. By reducing this parameter, the number of introduced inconsistent pairs increase and this helps to identify some observations are participating in many inconsistent pairs.

For instance, in 2009, by setting maximum feasible variability to four days, an observation was found that participated in five inconsistent pairs (Figure 4.8). This observation reports the first flowering of wood anemone on 28th of February, which is an early time. However, the comments about this sighting clarified that volunteer observed yellow anemone species that flower earlier than wood anemone.

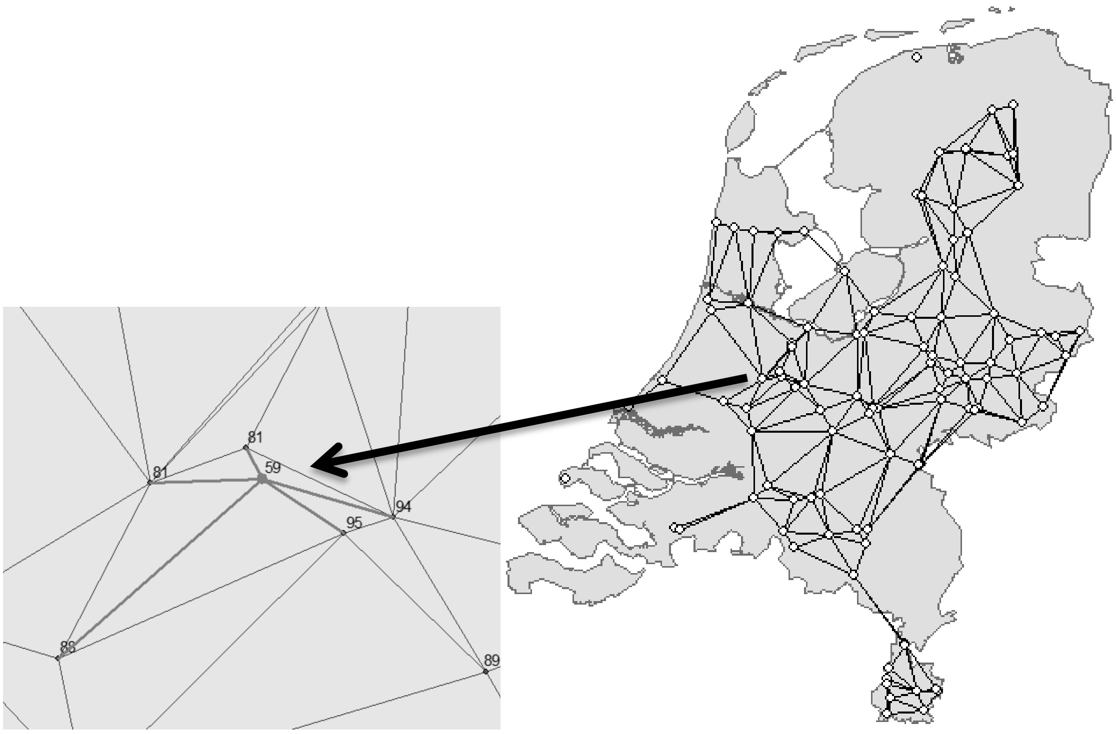



Figure 4.8. The first flowering of wood anemone in the fifty-ninth day of the year 2009 is introduced as an observation which is inconsistent with five of surrounding observations.

The selected proximity is another parameter that can affect the results since for smaller thresholds (e.g. 20 km) just remarkably close pairs can participate in constraint checking process. However, for datasets like the "Natuurkalender" observations are sparse and for thresholds smaller than 40 km many connections were lost. In this way, the enough numbers of observations do not participate in constraint checking increased.

## 4.4. Bias identification

This section shows the results of using marked point pattern analysis for identifying potential bias including spatial bias, temporal bias and urban-effect in the "Natuurkalender" observation. For this purpose, the wood anemone and swift and great tit are taken to illustrate how this analysis might be useful for informing such bias to users who are interested in phenological VGI.

### 4.4.1. Spatial bias

After making the decision about inconsistent observations, which means keeping or remove them, as described in sub-section 4.3.3, yearly observations of wood anemone were subjected to investigation of complete spatial randomness. In this research, inconsistent observations were considered in bias identification since there were no reference data to review them precisely.

Figure 4.9 depicted the prepared point pattern, which is surrounded by prepared observation windows (see sub-section 3.2.3), of wood anemone regarding to its possible habitat in 2008, which was chosen as an example, to present the result of spatial bias identification. In this analysis, it is assumed that the user is interested in using observations to study in an ecological unit area of 2500 and 625 square kilometres. Hence, the Quadrat counting test was conducted to test for complete spatial randomness regarding two quadrat sizes, 50km x 50km squares and 25km × 25km squares. The results of the test for each quadrat are presented in Table 4.9. The results (p-values) are calculated for a significance level of 0.001. This is because the number of observations in each year was low (with respect to the study area) and larger significance levels (e.g. 0.05) rejected the null hypothesis for all species.

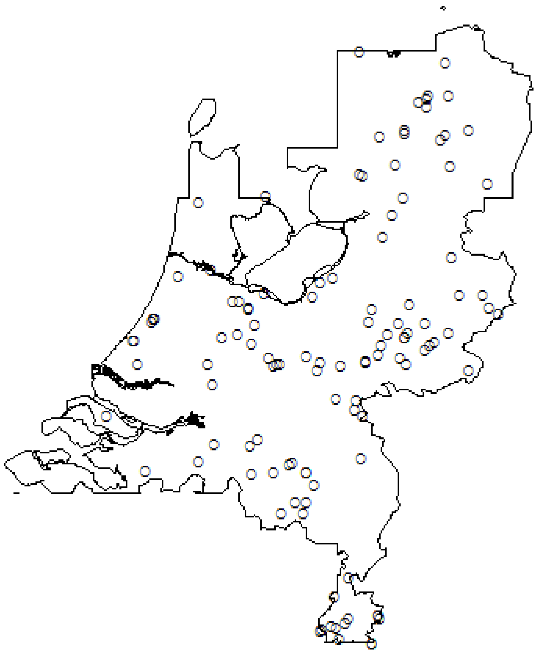


Figure 4.9. Prepared point pattern and observation window from wood anemone observations in 2008

Table 4.9. The results of quadrat counting test on wood anemone observations.

| | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|
| **50km X 50km squares** | 0.00001 | 0.005 | 0.00001 | 0.003 | 0.004 | 0.001 | 0.0008 |
| **25km X 25km squares** | 0.01 | 1.4 e-15 | 0.01 | 0.02 | 1.9 e-10 | 7.3 e-28 | 2.3 e-24 |

For 50km × 50km squares, the test rejects the uniform Poisson distribution in year 2003, 2004, 2005 and 2009. While, for years 2006, 2007 and 2008 the distribution of observation can be considered spatially random in the study area. While, for 25km × 25km quadrat, only years 2003, 2005 and 2006 are not involved with spatial bias.

To present the result of this examination in an interpretable way to users, the maps of empty space distances were produced for each year. They are presented in Figure 4.10 when the reference locations are the centre of 1km × 1km grid cells. In fact, these raster layers show in which part of the study area there is a lack of observations.

For instance, in 2005 which introduced the smallest p-value in quadrat counting test (when quadrats are 50km × 50km squares), the observations are most concentrated in the south and eastern part of the species habitat. During this year, areas coloured in yellow and orange are about 40km far from nearest volunteered observation.

These raster layers have different applicability. First, identify the area that involved in spatial bias either in the meanwhile or after collecting volunteered observations. Second, illustrate how the spatial distribution of the volunteered observations varies considerably from year to year. Thirdly, identify areas that during the years introducing spatial bias. For instance, it is notable to observe that the largest distances (yellow colours) are at the edges of the observation window.

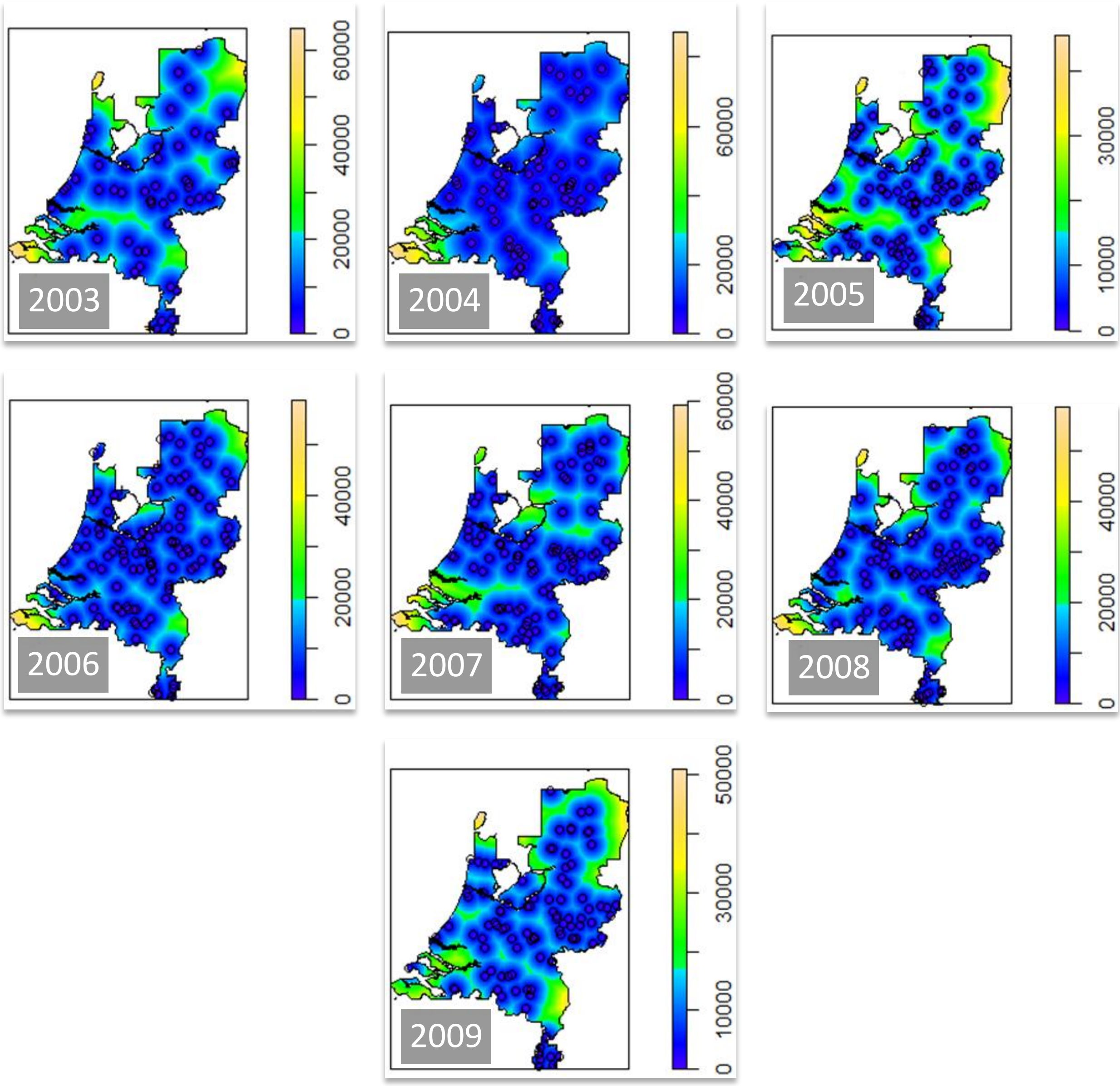



Figure 4.10. The empty space distance maps of wood anemone

#### 4.4.2. Temporal bias

For wood anemone, swift and great tit, the date of observations was converted to the weekend and non-weekend days (see sub-section 3.2.1). The expected proportion of observation done in weekend days is 0.28 because the chance of observing species in every day, in the week, is assumed equal. This expectation is tested by applying the Pearson's $\chi2$ test to check the goodness-of-fit, and the results (under significance level of 0.05) are depicted in Table 4.10 and Table 4.11.

Table 4.10. The results of the weekend-effect test for wood anemone (2003-2009)

| | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|
| **p-value** | 0.001 | 0.97 | 0.85 | 0.02 | 0.001 | 2.0e-5 | 0.001 |
| **Proportion of observation done on the weekend** | 0.41 | 0.28 | 0.29 | 0.37 | 0.42 | 0.46 | 0.43 |
| **Number of observations** | 67 | 83 | 105 | 123 | 117 | 118 | 109 |

Table 4.11. The results of the weekend-effect test for great tit (2003-2009)

| | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|
| **p-value** | 0.64 | 0.72 | 0.68 | 0.06 | 0.75 | 0.54 | 0.20 |
| **Proportion of observation done on the weekend** | 0.32 | 0.30 | 0.30 | 0.41 | 0.27 | 0.25 | 0.37 |
| **Number of observations in year** | 31 | 54 | 75 | 41 | 130 | 60 | 52 |

For wood anemone, except for 2004 and 2005, the weekend effect exists as predominant phenomenon. For instance, about half of the observations happened over the weekend in 2008 which in The Netherlands where many people do not live in a forest, is not surprising.
For great tit, there is no weekend effect in all years; however, the p-value (0.06) is to cross the significance level in 2006, and this means for such common birds the weekend effect is possible also. For example, by testing the weekend effect for swift, species that are also common in the urban area, the results showed that the weekend effect exists for the most years (5 years of 7 years), the results are listed in Table 4.12.

Table 4.12. The results of the weekend effect test for swift (2003-2009)

| | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|
| **p-value** | 0.10 | 0.22 | 0.00001 | 0.002 | 0.00009 | 1.1e-19 | 0.0001 |
| **Proportion of observation done on the weekend** | 0.22 | 0.32 | 0.40 | 0.19 | 0.38 | 0.55 | 0.40 |
| **Number of observations in year** | 132 | 245 | 265 | 227 | 302 | 239 | 202 |

#### 4.4.3. Urban-effect identification

In this, the result of examining urban-effect are presented for two species, wood anemone and swift, to realize to what extent the adopted method can work efficiently. For wood anemone, five possible land cover (cropland/natural vegetation mosaic, urban and built-up, cropland and mixed forests) that this species can be observed are distinguished from MODIS land cover datasets. The probability of observing this species in each of the five land cover is assumed equal. One of the selected land cover classes is the urban areas, which is assigned the probability 1/5, and consequently all other four land cover classes have the probability 4/5 for being a land cover. Except snow and ice land covers, all land cover classes are identified as possible land cover in which swift can be observed. Then, the probability of observing swift in urban land cover assigned 1/16 and for other land cover it assigned 15/16 are assigned to compare the distribution of "Urban_Nonurban" variable with Pearson's $\chi^2$ distribution. In this sense, the finesse of yearly observations of both species was tested by Pearson's $\chi 2$ test (under the assumption a significance level of 0.001). The results are listed in the Table 4.13 and Table 4.14.

Table 4.13. The results of urban-effect test for wood anemone observations (2003-2009)

| | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|---|
| **p-value** | 0.21 | 0.0006 | 0.34 | 0.00008 | 0.001 | 0.02 | 0.003 |
| **Proportion of observation done in the urban land cover** | 0.26 | 0.34 | 0.23 | 0.34 | 0.31 | 0.28 | 0.31 |
| **Number of observations** | 67 | 83 | 105 | 123 | 117 | 118 | 109 |

Table 4.14. The results of urban-effect test for swift observations (2003-2009)

| | **2003** | **2004** | **2005** | **2006** | **2007** | **2008** | **2009** |
|---|---|---|---|---|---|---|---|
| **p-value** | **p-value is near zero** | | | | | | |
| **Proportion of observation done in the urban land cover** | 0.543 | 0.444 | 0.506 | 0.491 | 0.512 | 0.489 | 0.457 |
| **Number of observations** | 132 | 245 | 265 | 227 | 302 | 239 | 202 |

As is expected, for swift, the volunteered observations are highly biased toward the urban area since in all years the results of the test reject an $\chi 2$ distribution as the null hypothesis. These results are not unusual since these species can move through the areas with urban land cover where the most volunteers are living, and then they can be seen by volunteers in this land cover type first. On the other hand, for wood anemone, the test only rejects a $\chi 2$ distribution for the years 2004 and 2006 (under significance level of 0.001). Since this species has a specific habitat, it is not expected to be observed mostly in urban areas, which this test confirmed also.

# 5. A CONCEPTUAL HUMAN-COMPUTER WORKFLOW

## 5.1. Introduction

Combining the power of machines and humans to solve problems is an appealing topic today. The integration of the speed and scalability of mechanical computation with the real comprehension of human computation can solve complicated computational problems (Law and Ahn, 2011). For an example, "FoldIt[2]" project attempts to predict the structure of a protein by taking advantage of computers' and humans' problem solving abilities.

In environmental monitoring, there has been an increase in human-computer network to collect and assess the quality of data. For instance, "eBird[3] " a real-time, online checklist program that engages a global network of volunteers to report bird observations (eBird, 2012). This program benefits from both human and computer algorithms to produce accurate estimates of species distributions (Sullivan et al., 2009) by a quality checking workflow that relies on active learning and feedbacks.

However, their filtering method relies on frequency of reporting an event or organism. For instance, Figure 5.1 depicts how the annual observations of Chipping Sparrow (a bird) were filtered in two counties, in New York. In this figure, the acceptable range of occurrence of this bird are displayed by dark bars, which extracted by accepting a range of the frequency distribution of historical stored observations in their databases.

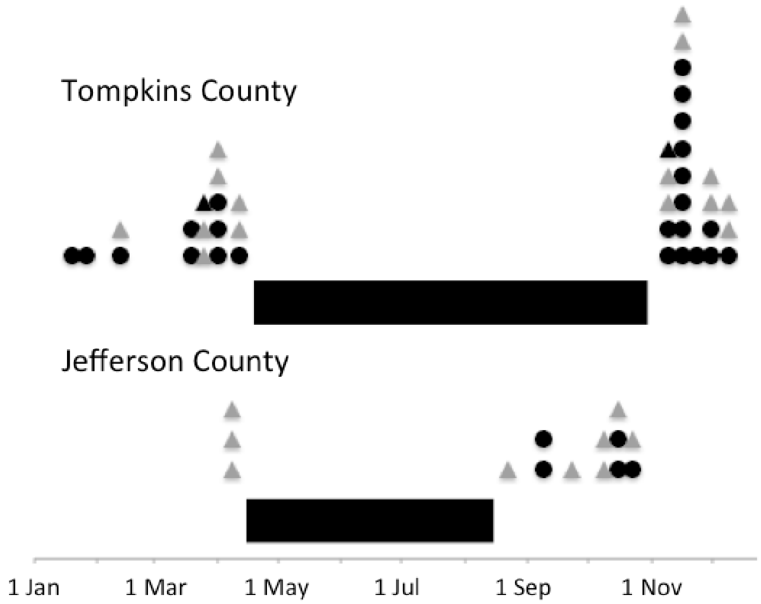



Figure 5.1. All observations that fall outside of the acceptable date range are filtered, which are done by either amateur volunteers (circles) or experienced volunteers (triangles) (Kelling et al., 2011).

The leading chapter aims at proposing a conceptual interactive workflow that can assist users in the evaluation of fitness of observations for the purpose of phenological monitoring VGI. In addition, it helps to provide feedback to the administrator of phenological monitoring VGI project and volunteers. This workflow is provided for phenological monitoring VGI, but it is adaptable to other types of monitoring VGI. In the following sections, the components of such a workflow are explored to understand their interactivity and requirements.

---

[2] www.fold.it
[3] www.ebird.org

## 5.2. Human components

To give a brief overview, a high level workflow is depicted in Figure 5.2. This workflow involves a number of different types of users and each of them has a specific role, with corresponding interactions. Then, it is necessary to determine each of these roles in the workflow.

The first user in the workflow is the volunteer, who observe a species' phenophase and records the location and date. They submit their observations via an online form on the website of the project (e.g. www.natuurkalender.nl) in which some optional and mandatory attribute fields exist (Figure 5.3). For the involvement, the website offers the possibility of seeing a map with all the sightings of current and previous years (per species) but also provides news items and nature predictions and other services.

The second user is the project administrator, who decides on the variety of species and guides volunteers to what and to how collect the data via the project websites. Additionally, they decide on the schema of the database (e.g. rational) that is stored by their servers.

The third user is the expert-user. The expert-users in this workflow are the scientists who want to use the phenological monitoring VGI in their studies. Currently, scientists can retrieve the data from the website of the volunteer-base monitoring program only by viewing the map of the sightings (same as the volunteer). Alternatively, they can also request to receive data from the project database.

The proposed workflow assumes that all three involved users are concerned about bias and inconsistency in phenological monitoring VGI. In this way, the following section explores what computer components address such quality issues and how they interact with their users.

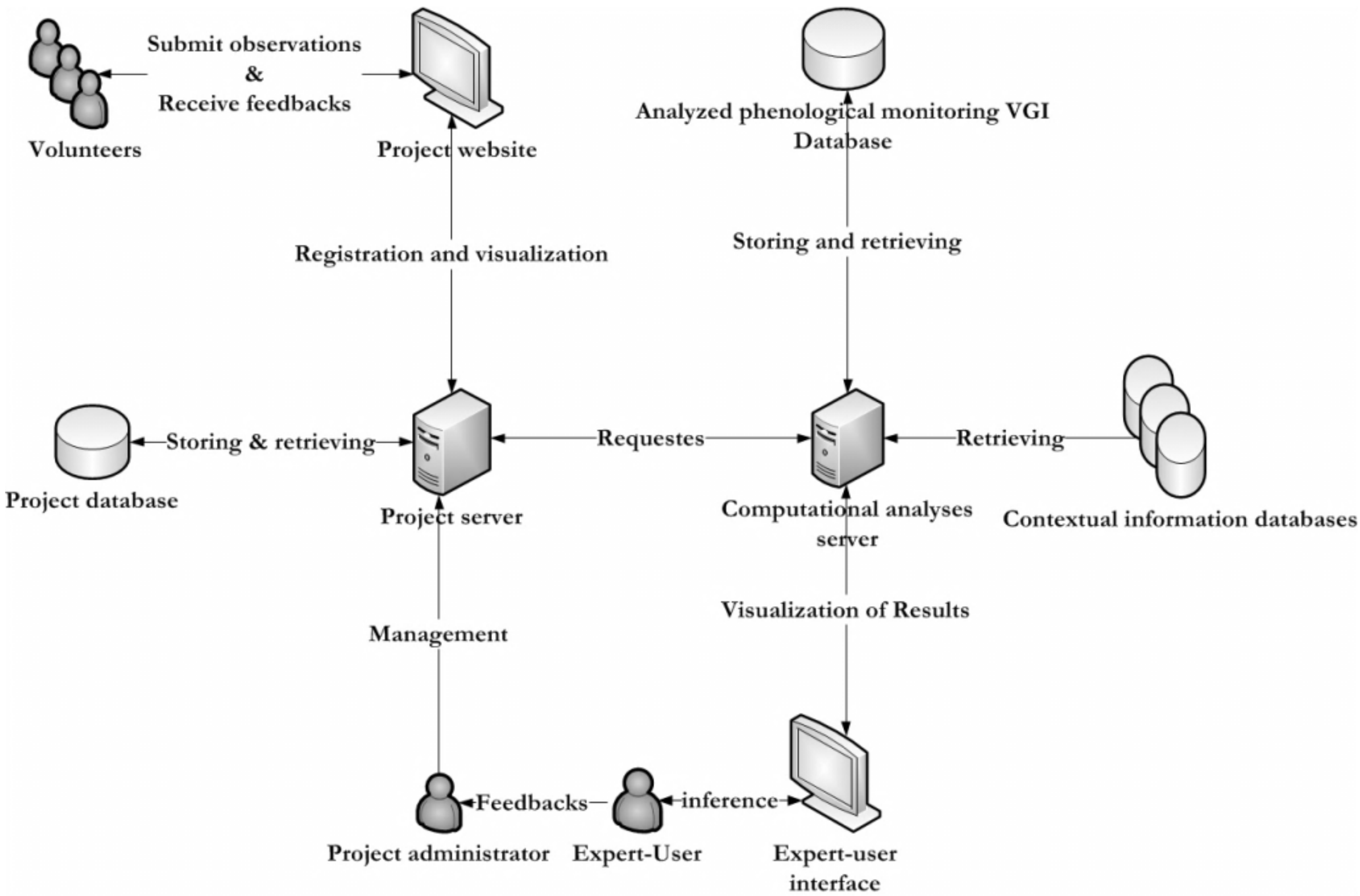



Figure 5.2. An interactive human-computer workflow to improve phenological monitoring VGI


Figure 5.3. The observation registration in the "Natuurkalender" project

## 5.3. Computational analyses component and expert-user interface

Using computers has a significant role in the proposed workflow since users are doing all data entering, storing, retrieving, analysing and visualizing of data they are using. In addition, for dealings with large-volume datasets, the computer components eases and speeds up these steps. In this way, they are applied in different phases of the workflow.

Firstly, after entering volunteers' observation via monitoring program website, data are stored, in the tabular format, and are visualized instantaneously for volunteers. For example, Figure 5.4 displaysy the web page in which one of the observations made by the author of this thesis, in the "Natuurkalender" project, is visualized (the largest point).

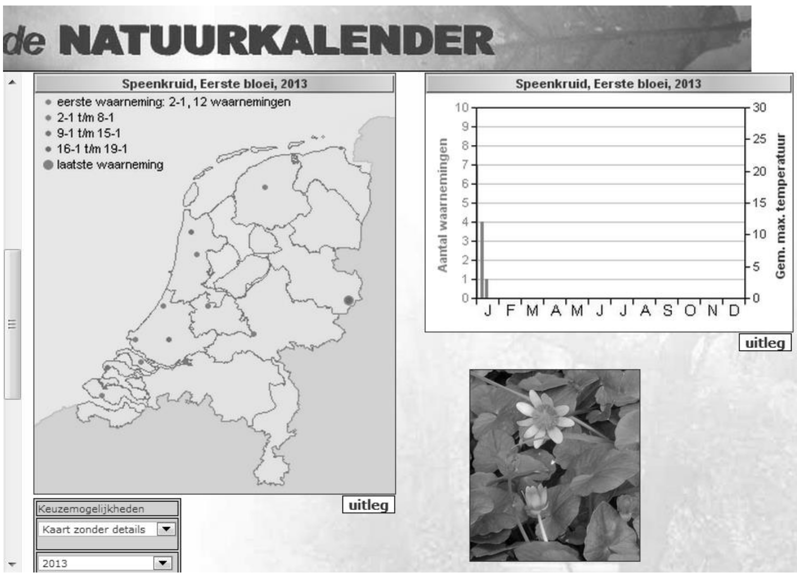



Figure 5.4. Instant visualization of registered observations in the "Natuurkalender" project

Secondly, based on a time interval that expert-user wants to work on, the computational analysis server tries to extract data from the maps on the monitoring web site or to request (if applicable) them directly from the project database. Moreover, data preparation method (see sub-section 3.2.1) can be implemented in this server, by retrieving and joining the contextual information from different databases (e.g. KNMI database).

This is attempted to consider two main tracks, in designed workflow that attempts at improving the quality of phenological VGI. For both, the computational analyses server plays a fundamental role. In one of the tracks, inconsistent and biased are identified and reflected to the expert users. These retrospective analyses are implemented in an interactive process which depends on expert-user requirements. It means that the expert-user selects a year (years), an observation proximity threshold which are interested in. Indeed, the interactivity of workflow comes from an interface (e.g. web application or computer software) which gives an exploring options to the expert-user. This interface translates the expert-user requirements in an understandable request format for the computational analysis server.

In addition, in such an interface, all information about inconsistent pairs can be queried from the result of data preparation step and be shown by click on it. Alternatively, the identification of spatial bias, temporal bias and urban-effect bias can be conducted by setups that users impose (e.g. the quadrat sizes). Additionally, the server can identify inconsistency regarding to land cover classification (e.g. urban/non-urban) or inform the volunteers' comment to evaluate inconsistent observations. In this way, expert-users are aware of the quality of data that they are to use.

Regarding to the result of interaction between expert user and computational analyses server, feedbacks from the expert user to project administrator can enhance the predictive or instantaneous performance of the monitoring project. For instance, it may happen by advising administrators to add an attribute fields to data submission forms that classify volunteers' observations to very early, normal and very late.

On another track, the workflow benefits computational analysis server to give immediate feedbacks to volunteers automatically. These feedbacks include the inconsistency checks based on highly correlated climate variable for newly registered observations as well as empty space distance maps. The feedbacks can be in the form of requests to the volunteers whom their observations are identified as inconsistent. In addition, meanwhile of data collection, updated empty space distance maps on the project website can motivate volunteers to areas involved in lack of observation (Kelling et al., 2011).

# 6. CONCLUSIONS AND RECOMMENDATION

## 6.1. Introduction

The main objective of this research was adopting analysis methods to identify potential inconsistency and bias in volunteered phenological observations besides proposing a conceptual human-computer workflow to inform about them in an interactive way. A range of analysis methods was implemented to show how contextual information, human perception and computer capability can be integrated to identify inconsistencies and bias.

The research showed how this can be achieved by a combination of exploratory data analysis, constraint satisfaction approach, marked point pattern analysis and statistical hypothesis testing methods. In addition, a conceptual workflow is proposed that showed how these methods can be integrated (into the Natuurkalender website) to improve the expertise of a range of users includes expert-users, project administrators and volunteers who are interested in phenological VGI. In the following sections, the conclusions achieved in each research step are described as well as the challenges were faced in this research. Moreover, some recommendations were made for future works.

## 6.2. Conclusions

In this research, the main objective was met with the accomplishment of the four research sub-objectives presented in the introduction chapter. This section clarifies how each sub-objective was achieved by answering to its corresponding research question:

- **How can contextual information be applied to support methods that check for inconsistency in volunteered phenological observations?**

The idea was to explore the distribution of phenophase time and its correlation with cumulative climate variables to select highly correlated variable, as well as, apply highly correlated variable as a constraint for inconsistency identification. For this purpose, exploratory data analysis based on histograms, scatter plots and Pearson's correlation test was applied, and it was able to discover the high correlations. In addition, linear regression was used to extract an applicable constraint.

For both plants and birds, the cumulative sum of daily evaporation was found highly correlated variable with the day of observations in the year (DoY); however, for plants, the correlations were higher than birds as it was expected. Additionally, for both observations done in the urban land cover and non-urban land cover, the implementation of methods showed exactly the same result with no significant difference to complete number of observations. As an initial conclusion, this research treaded to extracting constraints by considering complete observation in each year.

Linear regression modelled the difference in phenophase time with respect to the difference in cumulative sum of daily evaporation. The slope of the fitted line showed how the change in cumulative sum of daily evaporation can illustrate the difference in phenophase time. In general, the difference in birds' phenophase time was less sensitive to the difference in cumulative SDE (about half) than the difference in plants' phenophase time. With respect to this, it could be concluded that better contextual information should be found for using in the inconsistency identification of birds' observations.

In summary, the methods applied to answer this research question, explored and modelled the species' phenophase time regarding cumulative climate variable. The results were used to specify a constraint, a range function, in which consistent observation should confirm each other. This range function could be assumed as input in methods that are concerned about consistency identification. As the main conclusion, the exploring contextual information and finding highly correlated variable can be useful in inconsistency identification in different study area, as well as, different domain of monitoring VGI.

- **How can inconsistent volunteered observations be identified in phenological monitoring projects?**

The second sub - objective was to adopt a method to identify inconsistent observations in volunteered phenological observations. For this purpose, a constraint satisfaction approach was used for both fixed and moving species. This approach benefited from contextual constraints extracted from preceding research sub-objective as well as proximity of observation which was defined by constructing a graph through yearly observations.

The approach was found to be sensitive to the distance and the feasible variation in species' phenophase time (under the same condition). According to volunteers' comment about observations, the constraint satisfaction approach could identify most of the inconsistent observations for plants. Nevertheless, this approach introduced more observation as the inconsistent in the case of birds. For birds, the results could not critically evaluate since no reference information (i.e. reliable information that shows the exact time of phenophase or comments from volunteers that show inconsistency or mistake about observations).

For wood anemone, the large number of introduced inconsistent observations related to the years that there has been mild spring like 2008 (nature has produced blooms about one month earlier) which make sense. However, there was not a logical pattern for the number of inconsistent observations for swift observations through 2003 to 2009.

In short, the constraint satisfaction approach used in this section identified the observations are sighted earlier or later than expected in each year by linking the difference in DoY to the difference in cumulative SDE for proximate observations. The approach showed acceptable results for plants while for the moving species like birds, the efficiency of the approach were not clear. This is because the number of identified inconsistency was large and there was no evidence to assess them. This implied that the mobility of species leads to introducing uncertainty about inconsistency which was expected to the smaller correlation coefficients for swift. Then, the approach should be further validated about moving species.

- **How can potential spatial and temporal bias in volunteered phenological observations be identified and be informed?**

In the third step of this research, a marked point pattern analysis was applied to identify potential spatial bias, temporal bias and urban-effect in three representative species in the "Natuurkalender" datasets, wood anemone, swift and great tit.

Marked point patterns were created from yearly observations of selected species. The result of the quadrat counting test, which was conducted for wood anemone, showed that the complete spatial randomness is not an absolute concept. The main conclusion of this examination was that occurrence of spatial bias completely depends on the scale of studies that want to use phenological VGI. For one study distribution of observation can be assumed complete spatial random while for another cannot be assumed.

On the other hand, the empty space distance map found as an informative product about spatial bias for users who are interested in phenological monitoring VGI. It could visualise both the areas that have potential to bias observation and the areas in which there is a lack of observations. As a conclusion, these maps were found more useful for volunteers than expert-users.

The urban-effect test applied on representative species, wood anemone and swift. For swift in all years there was an urban effect which is not surprising since this species is frequently seen in the urban environment that most of the volunteers live. While, in most of years the result of the test did not show urban-effect for wood anemone. The natural habitat of wood anemone normally is outside of the urban area then it is not a surprising result. From these results, it could be concluded that the designed test can efficiently reveal the bias may exist toward observing in the urban area for species are common in the urban environment.

Finally, the weekend-effect test was conducted for wood anemone, swift and great tit to identify potential temporal bias in the observations. The results showed that there was no weekend effect for great tit, which makes sense since it is a conspicuous species through the everyday life and the most of the public know it. The strange results related to swift which showed the weekend effect as the predominant phenomenon from year 2003 to 2009. This species is also a common species that exist in the urban area and everyday life, but for 5 years of 7 years there was a weekend effect. Since the swift is flying at an altitude higher than the great tit, the volunteers may be more aware of it as they are less busy at weekends. The general conclusion is that marked point pattern analysis gives the facility about identifying spatial bias, temporal bias and urban-effect in phenological VGI together.

- **What are the components, requirements and interrelationships in a workflow to inform and improve the quality of volunteered phenological observations in an interactive way?**

In chapter 5, a conceptual workflow, which applies both human interpretation and computer capability, was proposed to inform and improve the inconsistency and bias in monitoring VGI. Three types of users, who are interested in phenological VGI and their requirements were introduced. The users included volunteers, project administrator and expert-users. In addition to users, the computational analysis server and expert-user interface was determined as computer parts can facilitate the production and communication of information about bias and inconsistency.

Two different types of use for inconsistency checks were identified, for analysis by the expert users and for fast feedback to volunteers who entered an inconsistent observation. The method proposed for finding inconsistencies, looks very promising for the expert user, but may not be implementable for feedback to the volunteers. This is because, for first observations in each year, there is no enough number of observations to both extracting the contextual constraint and constructing constraint graph.

Two methods were tested for identifying spatial bias, the quadrat counts and the empty space maps. These two differ in the way the results are presented. The quadrat counts like to a quantification of the bias, which is more suitable for expert-users, however, the empty space maps visualise the bias in such a way that even the non-expert volunteer can easily understand the result.

For temporal bias and urban-effect, statistical hypothesis testing was proposed in the computational analysis server. The "yes" or "no" are answers to users are concern about this type of bias. Regarding the scale (spatially or temporally) of user requirements, the computation analysis server of workflow attempted to produce statistical and visual information which can improve the expertise of volunteers, enhance the awareness of project administrator and phenologist.

## 6.3. Research Challenges

In this section, the challenges faced in this research are discussed. First, the number of yearly observations for selected species was low (e.g. wood anemone observations in 2003 there were just 67 observations) that can lead to uncertainty about extracting constraints by linear regression. As the result, some of the introduced pairs as inconsistent in such years may not be surely inconsistent, and they should be review precisely by experts.

Second, the results of the constraint satisfaction approach can be enhanced if more yearly observations were available. In this way, the method can be implemented for smaller proximity and more observations can participate in constraint checking. As the result, the accuracy of results will be increased. However the minimum required number is not explicit to this research yet.

Third, for some species in the "Natuurkalender" project (e.g. butterflies); there was a high variation in distribution of phenophase timing (DoY) through the year. This prevents to extract a reliable constraint from modelling (Figure 6.1). As the result, the constraint satisfaction approach introduces a large number of pairs as inconsistent observations.

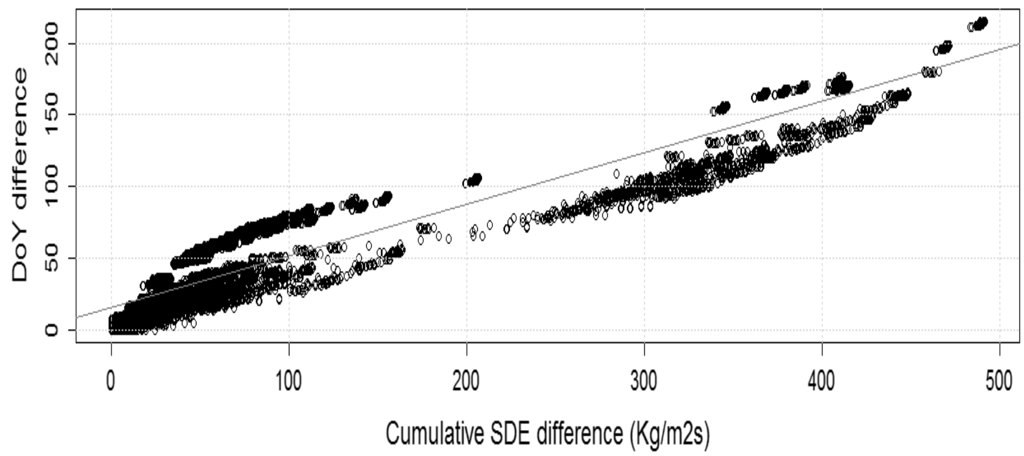



Figure 6.1. An extracted unreliable constraint for brimstone butterfly (2008)

Finally, because the definitions of the classes in the MODIS land cover datasets are not fully matching with habitats of species, there is uncertainty about the results of the urban-effect test. For example, 0.22 of observation of wood anemone in 2009 are done in the cropland which normally is impossible, in reality, since the natural habitat of this specie is woodlands.

In The Netherlands, land cover is highly various, and it is not surprising to find more than one land cover type in the area covered by a pixel in MODIS land cover datasets. This indicates that the resolution of the MODIS land cover product is coarse for this research. Thus, there is an uncertainty about the real land cover of observations. For instance Figure 6.2 shows that many observations are done in the boundaries of cities where the risk of misclassification of land cover is high. Then, it may lead to the classification of a mixed forest as the urban and built-up land cover.

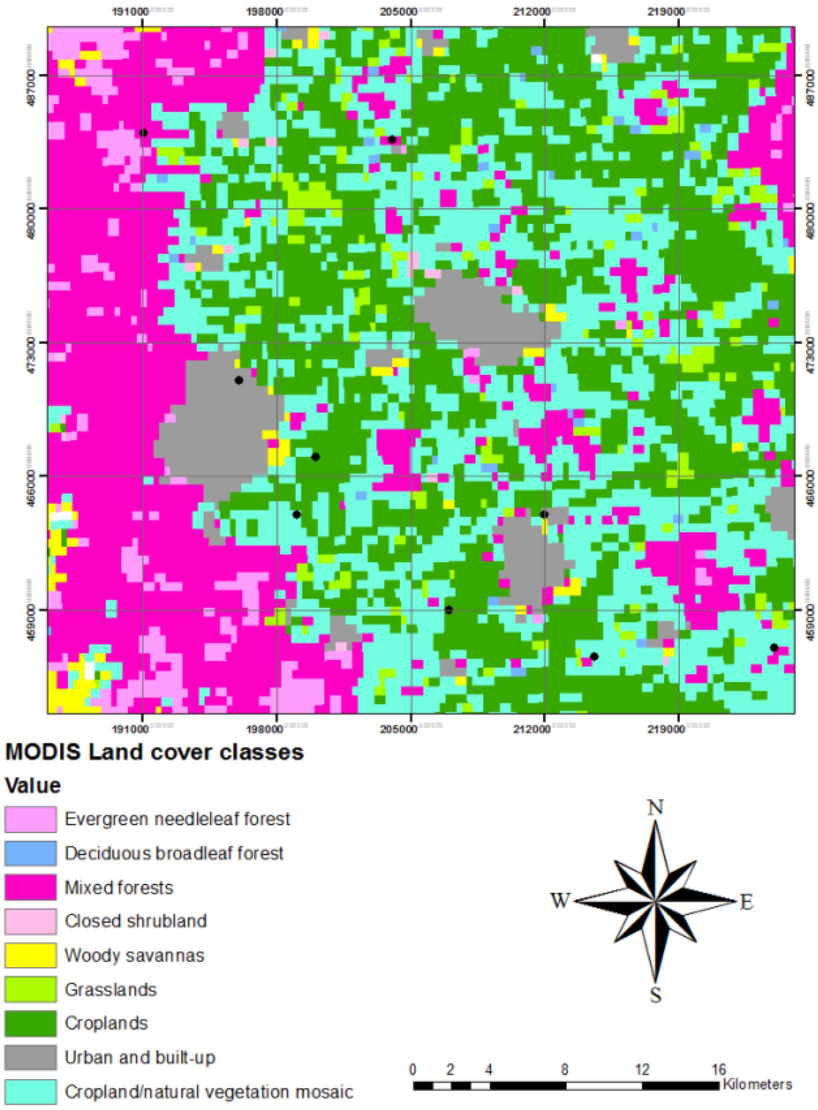



Figure 6.2. A sub-set of the MODIS land cover map (in The Netherlands) in which volunteers' observations plotted with black dots

## 6.4. Recommendations

With respect to conclusions and challenges, the following list presents some recommendations for further research and development.

- **Reference observations:** for critically assess the result of a constraint satisfaction approach, reference dataset is needed. These datasets can be provided by phenologist or be constructed synthetically. Providing and quantification the efficiency of proposed method can be a good idea to continue the work.
- **Classification of empty space distances:** the raster layer of empty space distance can be classified with respect to the distance. Classification products can be published online and be updated meanwhile by observation collection time to hint volunteers about the next location for observation.

- **The minimum required number of observations in constraint satisfaction approach:** as one of the input parameters for constraint satisfaction approach, the minimum required number of observations in each year found critical. Regarding to limit of time, the issue is not allow to be investigated. This implies a deeper field into the constraint satisfaction approach for phenological VGI itself.
- **Integrated software:** in this research, implementation of methods is done in different platform. Integration all methods as software for identifying inconsistency and bias in VGI can be a good idea for further works.
- **Test constrain satisfaction approach for other domain:** this research recommends for using the constraint satisfaction approach in other monitoring VGI to evaluate its power in consistency identification.

# LIST OF REFERENCES


Baddeley, A. 2008. Analysing spatial point patterns in R. Technical report, CSIRO, 2010. Version 4. Available at www. csiro. au/resources/pf16h. html.

Bessiere, C. 2006. Constraint propagation. *Foundations of Artificial Intelligence,* 2**,** 29-83.

Betancourt, J. L., Schwartz, M. D., Breshears, D. D., Brewer, C. A., Frazer, G., Gross, J. E., Mazer, S. J., Reed, B. C. & Wilson, B. E. 2007. Evolving plans for the USA National Phenology Network. *Eos, Transactions American Geophysical Union,* 88**,** 211.

Bishr, M. & Mantelas, L. 2008. A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal,* 72**,** 229-237.

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K. & Mace, G. M. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS biology,* 8**,** e1000385.

Brando, C., Bucher, B. & Abadie, N. 2011. Specifications for User Generated Spatial Content. *Advancing Geoinformation Science for a Changing World***,** 479-495.

Catlin-Groves, C. L. 2012. The Citizen Science Landscape: From Volunteers to Citizen Sensors and Beyond. *International Journal of Zoology,* 2012.

Cohn, J. P. 2008. Citizen science: Can volunteers do real research? *BioScience,* 58**,** 192-197.

CPP. 2013. *California Phenology Project* [Online]. Available: http://www.usanpn.org/cpp/node/8 2013].

De Beurs, K. M. & Henebry, G. M. 2004. Land surface phenology, climatic variation, and institutional change: Analyzing agricultural land cover change in Kazakhstan. *Remote Sensing of Environment,* 89**,** 497-509.

Delaney, D. G., Sperling, C. D., Adams, C. S. & Leung, B. 2008. Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions,* 10**,** 117-128.

Deparday, V. 2010. Enhancing Volunteered Geographical Information (VGI) Visualization with Open Source Web-Based Software.

Devictor, V., Whittaker, R. J. & Beltrame, C. 2010. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and distributions,* 16**,** 354-362.

Dickinson, J., Zuckerberg, B. & Bonter, D. N. 2010. Citizen science as an ecological research tool: challenges and benefits. *Annual review of ecology, evolution, and systematics,* 41.

eBird. 2012. *Global tools for birders, critical data for science — eBird* [Online]. Available: http://ebird.org/content/ebird/about 2012].

Fargier, H., Lang, J., Martin-Clouaire, R. & Schiex, T. A constraint satisfaction framework for decision under uncertainty. Proc. of the 11th Int. Conf. on Uncertainty in Artificial Intelligence, 1995. Citeseer.

Ferrer, X., Carrascal, L. M., Gordo, Ó. & Pino, J. 2006. Bias in avian sampling effort due to human preferences: an analysis with Catalonian birds (1900–2002). *Ardeola,* 53**,** 213-227.

Flanagin, A. J. & Metzger, M. J. 2008. The credibility of volunteered geographic information. *GeoJournal,* 72**,** 137-148.

Galton, F. 1890. Kinship and correlation. *The North American Review,* 150**,** 419-431.

Gordo, O., Sanz, J. J. & Lobo, J. M. 2007. Environmental and geographical constraints on common swift and barn swallow spring arrival patterns throughout the Iberian Peninsula. *Journal of biogeography,* 34**,** 1065-1076.

Haight, F. A. & Haight, F. A. 1967. Handbook of the Poisson distribution. *In:* UR, W. (ed.) *Publications in operations research.* New York: Wiley.

Hudson, I. L. & Keatley, M. R. 2009. *Phenological research: methods for environmental and climate change analysis*, Springer.

Hudson, J. C. & Fowler, P. M. 1966. *The concept of Pattern in Geography*, Department of Geography, University of Iowa.

Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. 2008. *Statistical analysis and modelling of spatial point patterns*, Wiley-Interscience.

Kanevski, M., Pozdnoukhov, A. & Timonin, V. 2009. *Machine Learning for Spatial Environmental Data: theory, applications and software*, EPFL Press.

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W. K., Yu, J., Damoulas, T. & Gomes, C. 2011. eBird: A Human/Computer Learning Network for Biodiversity Conservation and Research.

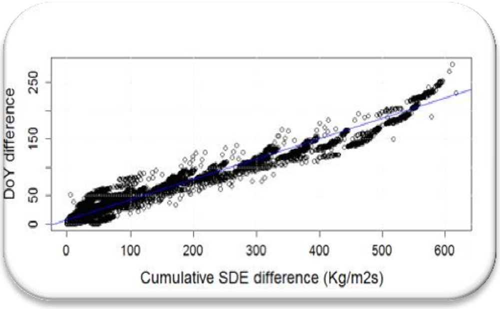
KNMI. 2004. *PRISM standard_name tables* [Online]. Available: http://www.knmi.nl/~velthove/PRISM/CF/PRISM_standard_names_V1.3.html [Accessed November 23 2012].
KNMI. 2012a. *KNMI About KNMI* [Online]. Available: http://www.knmi.nl/about_knmi/ 2012].
KNMI. 2012b. *KNMI Data Centre* [Online]. Available: https://data.knmi.nl/portal-webapp/KNMI-Datacentrum.html#location=eu 2012].
Law, B., Mackowski, C., Schoer, L. & Tweedie, T. 2000. Flowering phenology of myrtaceous trees and their relation to climatic, environmental and disturbance variables in northern New South Wales. *Austral Ecology,* 25**,** 160-178.
Law, E. & Ahn, L. 2011. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning,* 5**,** 1-121.
MacEachren, A. M. & Kraak, M. J. 2001. Research challenges in geovisualization. *Cartography and Geographic Information Science,* 28**,** 3-12.
Mackworth, A. K. 1977. Consistency in networks of relations. *Artificial intelligence,* 8**,** 99-118.
NASA. 2012. *MCD12Q1 | LP DAAC :: ASTER and MODIS Land Data Products and Services* [Online]. Available: https://lpdaac.usgs.gov/products/modis_products_table/mcd12q1.
Natuurkalender. 2012a. *The Nature* [Online]. Available: http://www.natuurkalender.nl/doelstelling.asp [Accessed November 12 2012].
Natuurkalender. 2012b. *The Nature* [Online]. Department of Environmental Systems Analysis Group , Wageningen UR. Available: http://www.natuurkalender.nl/partners.asp [Accessed December 22 2012].
Nizinski, J. J. & Saugier, B. 1988. A model of leaf budding and development for a mature Quercus forest. *Journal of Applied Ecology***,** 643-652.
Renz, J. & Nebel, B. 2007. Qualitative Spatial Reasoning Using Constraint Calculi. *In:* Aiello, M., Pratt-Hartmann, I. & Benthem, J. (eds.) *Handbook of spatial logics.* Springer Netherlands.
Rowlingson, B. S. & Diggle, P. J. 1993. Splancs: spatial point pattern analysis code in S-Plus. *Computers & Geosciences,* 19**,** 627-655.
Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M. & Edwards, D. D. 1995. *Artificial intelligence: a modern approach*, Prentice hall Englewood Cliffs, NJ.
Rutishauser, T. 2003. *Cherry tree phenology: Interdisciplinary analyses of phenological observations of the cherry tree in the extended Swiss plateau region and their relation to climate change.*
Schlieder, C. & Yanenko, O. Spatio-temporal proximity and social distance: a confirmation framework for social reporting. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, 2010. ACM, 60-67.
Shen, F. 2012. *Spatial and temporal analysis of volunteered phenological observations.* MSc, ITC,Faculty of Geo-Information Science and Earth Observation of the University of Twente.
Soortenbank. 2012. *SoortenBank.nl : Over deze website* [Online]. Available: http://www.soortenbank.nl/ 2012].
Sparks, T. H., Huber, K. & Tryjanowski, P. 2008. Something for the weekend? Examining the bias in avian phenological recording. *International Journal of Biometeorology,* 52**,** 505-510.
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**,** 2282-2292.
Tsang, E. 1993. *Foundations of constraint satisfaction*, Academic press London.
Tukey, J. W. 1977. Exploratory data analysis. *Reading, MA.*
Tulloch, D. L. 2008. Is VGI participation? From vernal pools to video games. *GeoJournal,* 72**,** 161-171.
Voříšek, P. 2008. *A best practice guide for wild bird monitoring schemes*, Pan-European Common Bird Monitoring Scheme (PECMBS).
Wageningen. 2012. *Nature's Calendar: The Dutch phenological network - Wageningen UR* [Online]. Wageningen UR. Available: http://www.wageningenur.nl/en/show/Natures-Calendar-The-Dutch-phenological-network.html.
Wiggins, A. & Crowston, K. From conservation to crowdsourcing: A typology of citizen science. System Sciences (HICSS), 2011 44th Hawaii International Conference on, 2011. IEEE, 1-10.
Yanenko, O. & Schlieder, C. 2012. Enhancing the Quality of Volunteered Geographic Information: A Constraint-Based Approach. *Bridging the Geographic Information Sciences***,** 429-446.
Zhang, X., Tarpley, D. & Sullivan, J. T. 2007. Diverse responses of vegetation phenology to a warming climate. *Geophysical Research Letters,* 34**,** L19405.

Zmihorski, M., Sparks, T. H. & Tryjanowski, P. 2012. The Weekend Bias in Recording Rare Birds: Mechanisms and Consequencess. *Acta Ornithologica,* 47**,** 87-94.

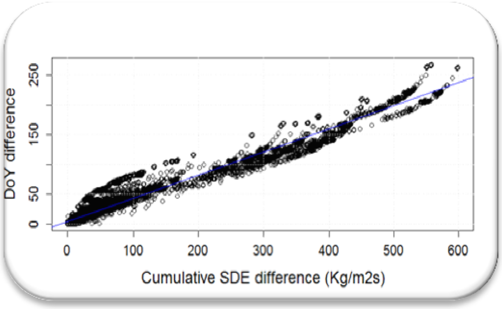
# APPENDICES

## Appendix A

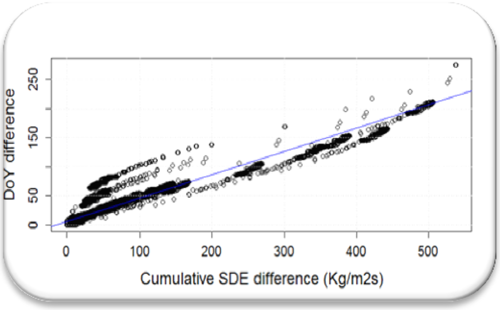
(1) Linear regression modelling for brimstone butterfly through 7 years (2003 to 2009)



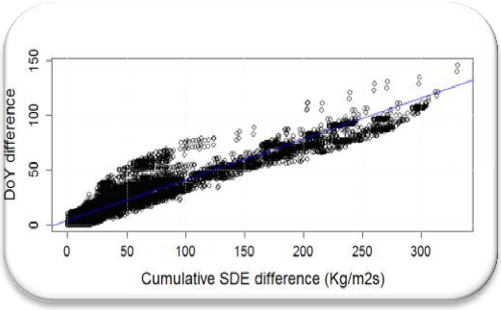

2003



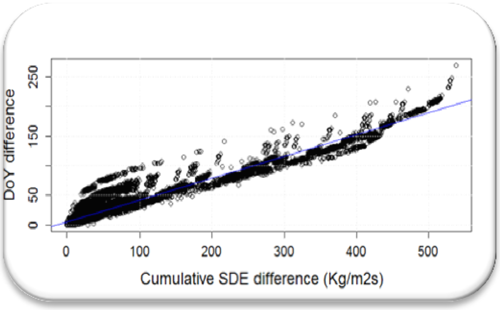

2004



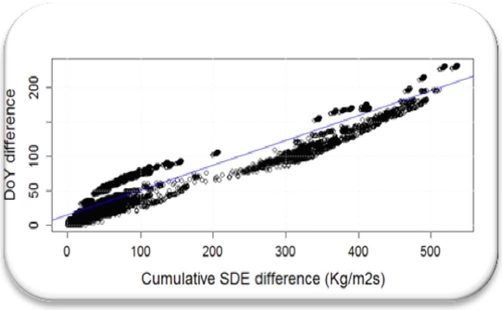

2005



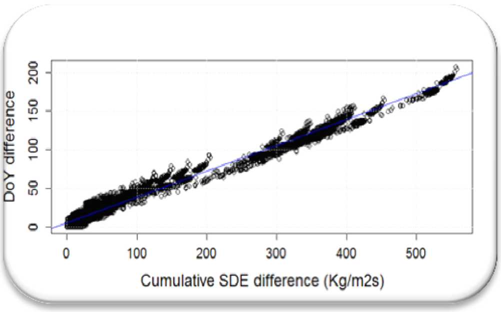

2006




2007

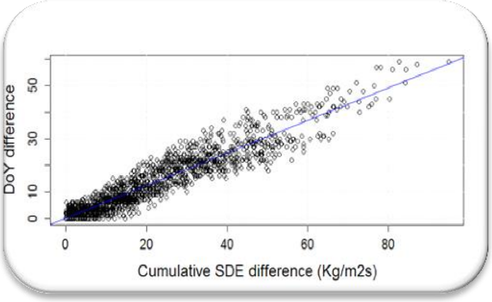



2008

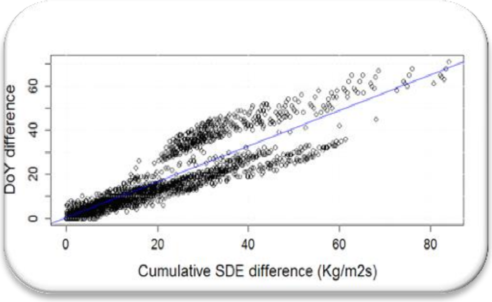



2009

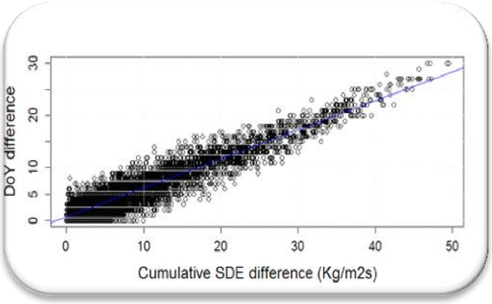
(2) Linear regression modelling for wood anemone through 7 years (2003 to 2009)



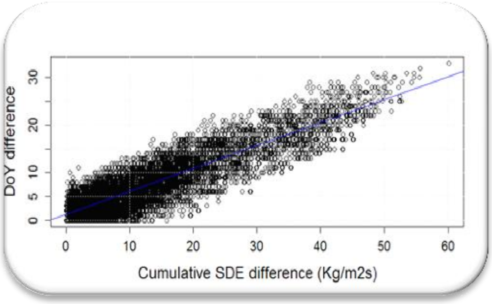

2003



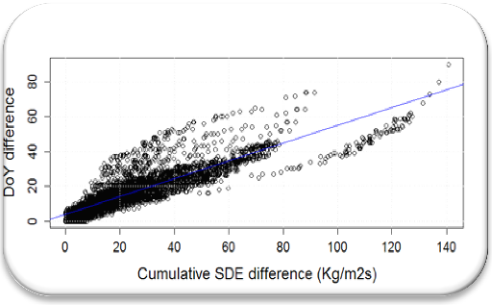

2004



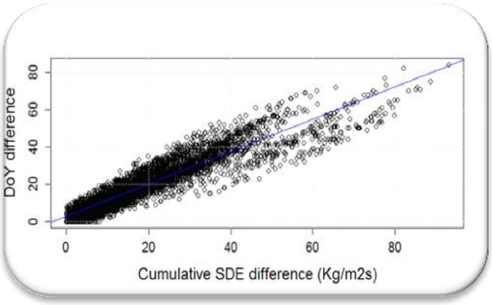

2005



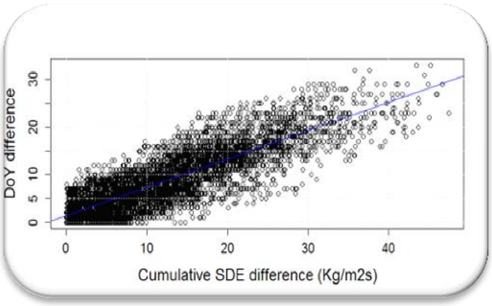

2006




2007

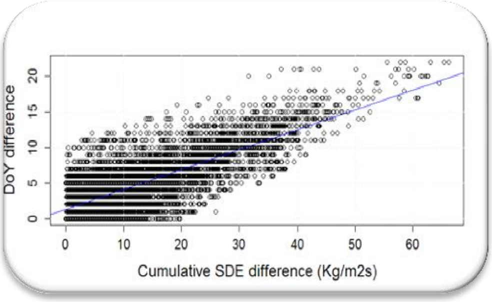



2008

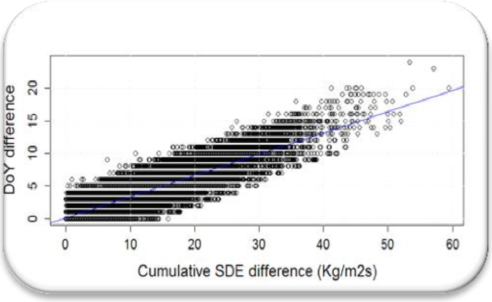



2009

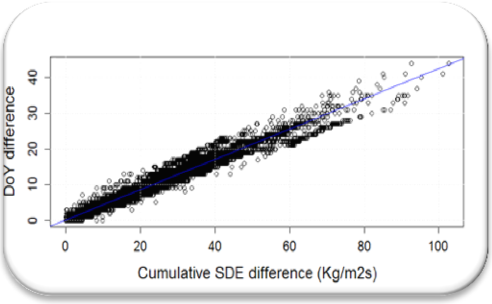
(3) Linear regression modelling for swift through 7 years (2003 to 2009)



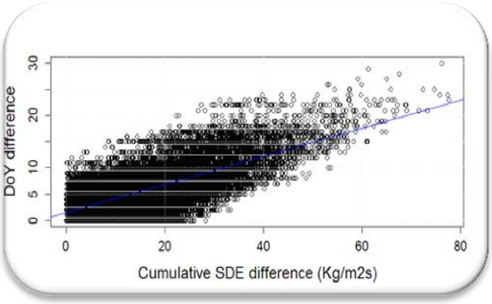

2003



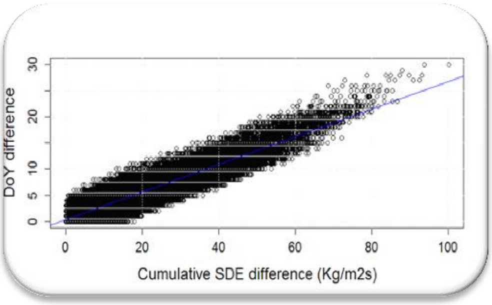

2004



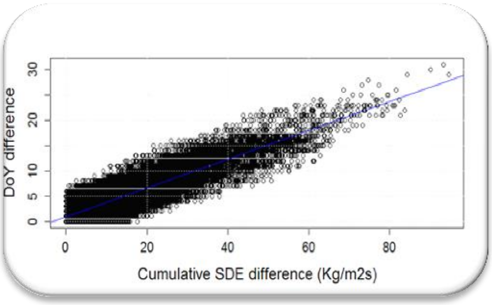

2005



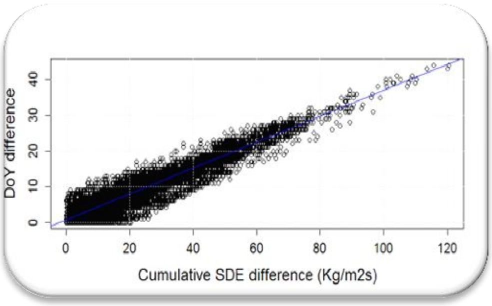

2006




2007

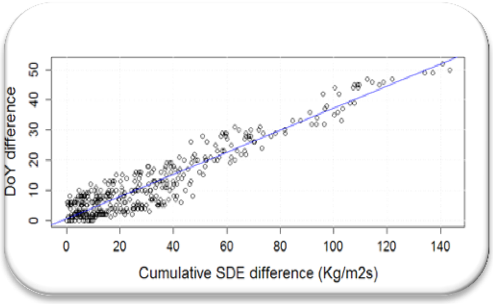



2008

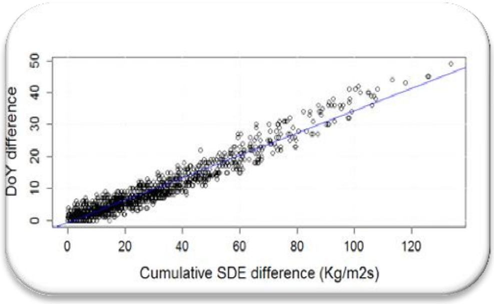



2009

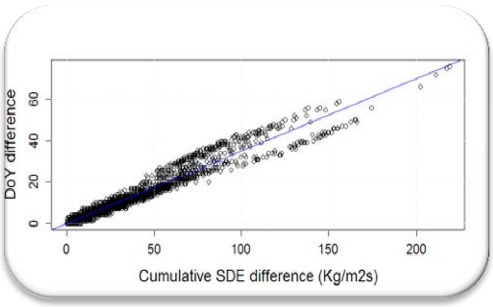
(4) Linear regression modelling for great tit through 7 years (2003 to 2009)



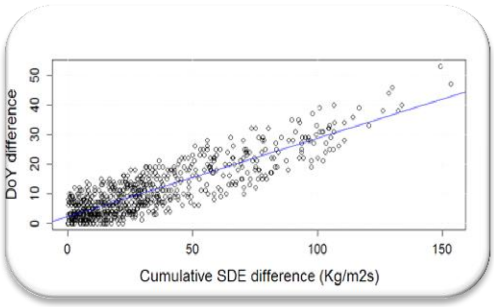

2003



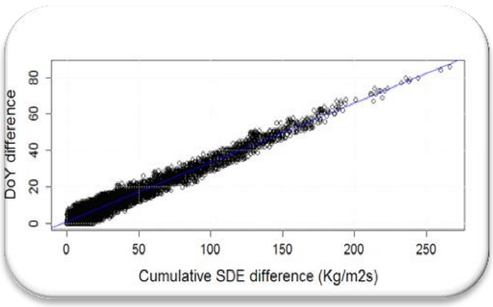

2004



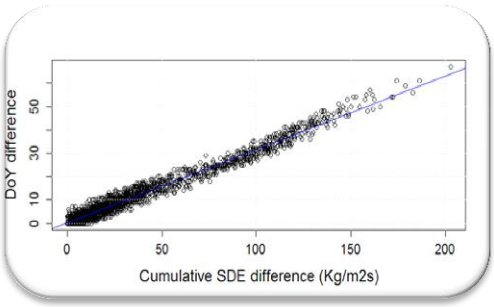

2005



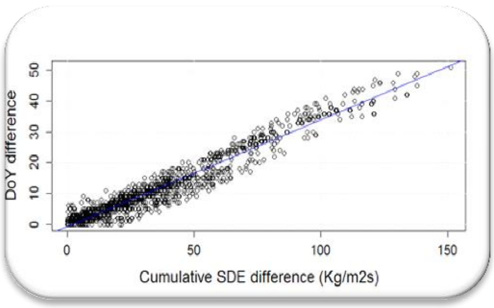

2006




2007




2008




2009

## Appendix B

R scripts used

```
##### importing "spatstat" and "maptools" modules #####
library(maptools)
library(spatstat)

##### Species selection #####
mydata = unique(read.csv(file.choose()))

##### Defining Study area boundary #####
winn = as(readShapeSpatial(file.choose()),"owin")

##### Selecting observations for individual year ######
mydata.individual.year <- which(mydata$Jaar == 2004)
sort(mydata$Jaar[mydata.individual.year])
individual.year <- mydata[mydata.individual.year, ]
individual.year[order(individual.year$nr), ]

##### Histogram of the day of observations in the year #####
hist(individual.year$DOY,xlab="Phenophase time (DoY)",ylim=c(0,50),breaks=seq(1, 365, by = 1),col =
"lightblue",cex.lab=1.5, cex.axis=1.25)
qqnorm(individual.year$DOY,cex.lab=1.5, cex.axis=1.25)
qqline(individual.year$DOY,cex.lab=1.5, cex.axis=1.25)

##### Creating marked point pattern #####
nature = ppp(individual.year$x, individual.year$y, window = winn, unitname=c("metre","metres"), marks
= individual.year$DOY,check = TRUE)
summary(nature)
summary(duplicated(nature))
plot(nature, cols="blue")
plot(hist(marks(nature), breaks = seq(0,16, by=1)))

##### Classification of observation to urban/non-urban and weekend/non-weekend #####
nature.urban = ppp(urban$x, urban$y, window = winn, unitname=c("metre","metres"), marks =
urban$Land_cover,check = TRUE)
plot(nature.urban)
hist(urban$DOY,xlab="Julian day",breaks=seq(1, 365, by = 1))
nature.nonurban = ppp(nonurban$x, nonurban$y, window = winn, unitname=c("metre","metres"), marks
= nonurban$Land_cover,check = TRUE)
plot(nature.nonurban)
hist(nonurban$DOY,xlab="Julian day",breaks=seq(1, 365, by = 1))
individual.year.weekend <- which(individual.year$DW == 'Weekend')
sort(individual.year$DW[individual.year.weekend])
weekend <- individual.year[individual.year.weekend, ]
weekend[order(weekend$nr), ]
nature.weekend = ppp(weekend$x, weekend$y, window = winn, unitname=c("metre","metres"), marks =
weekend$DW,check = TRUE)
```

```
plot(nature.weekend)
hist(weekend$Raster_value,xlab="Land cover class")
individual.year.weekday <- which(individual.year$DW == 'Weekday')
sort(individual.year$DW[individual.year.weekday])
weekday <- individual.year[individual.year.weekday, ]
weekday[order(weekday$nr), ]
nature.weekday = ppp(weekday$x, weekday$y, window = winn, unitname=c("metre","metres"), marks =
weekday$DW,check = TRUE)
plot(nature.weekday)
hist(weekday$Raster_value,xlab="Land cover class")

##### Scatterplot of Variables and correlations ########
##### All land covers #####
plot(individual.year$T0_sum,individual.year$DOY,ylab="DoY",xlab="Cumulative ADT (C)",cex.lab=1.5,
cex.axis=1.25)
grid()
plot(individual.year$precip_sum,individual.year$DOY,ylab="DoY",xlab="Cumulative SDP
(Kg/m2s)",cex.lab=1.5, cex.axis=1.25)
grid()
plot(individual.year$EV_sum,individual.year$DOY,ylab="DoY",xlab="Cumulative SDE
(Kg/m2s)",cex.lab=1.5, cex.axis=1.25)
grid()
cor.test(individual.year$DOY,individual.year$T0_sum)
cor.test(individual.year$DOY,individual.year$precip_sum)
cor.test(individual.year$DOY,individual.year$EV_sum)

##### Urban land cover #####
plot(urban$T0_sum,urban$DOY,ylab="DoY",xlab="Cumulative ADT (C)")
grid()
plot(urban$precip_sum,urban$DOY,ylab="DoY",xlab="Cumulative SDP (Kg/m2s)")
grid()
plot(urban$EV_sum,urban$DOY,ylab="DoY",xlab="Cumulative SDE (Kg/m2s)")
grid()
cor.test(urban$DOY,urban$T0_sum)
cor.test(urban$DOY,urban$precip_sum)
cor.test(urban$DOY,urban$EV_sum)

##### Non-urban land cover#####
plot(nonurban$T0_sum,nonurban$DOY,ylab="DoY",xlab="Cumulative ADT (C)")
grid()
plot(nonurban$precip_sum,nonurban$DOY,ylab="DoY",xlab="Cumulative SDP (Kg/m2s)")
grid()
plot(nonurban$EV_sum,nonurban$DOY,ylab="DoY",xlab="Cumulative SDE (Kg/m2s)")
grid()
cor.test(nonurban$DOY,nonurban$T0_sum)
cor.test(nonurban$DOY,nonurban$precip_sum)
cor.test(nonurban$DOY,nonurban$EV_sum)
```

```
##### Linear regression #####
data_type=individual.year
highly_correlated_variable = data_type$EV_sum
dim_data_type=dim(data_type)
dim_data_type[1]
Del_JD = matrix(nrow = dim_data_type[1], ncol = dim_data_type[1])
Del_highly_correlated_variable = matrix(nrow = dim_data_type[1], ncol = dim_data_type[1])
for(i in 1:dim_data_type[1]){
  for(j in 1:dim_data_type[1]){
    Del_JD[i,j] = abs(data_type$DOY[i]-data_type$DOY[j])
    Del_highly_correlated_variable[i,j] = abs(highly_correlated_variable[i]-highly_correlated_variable[j])
  }
}
Del = matrix(nrow = (dim_data_type[1]*dim_data_type[1]), ncol = 2)
u = 1
for (i in 1:dim_data_type[1]){
  for (j in 1:dim_data_type[1]){
    Del[u,1] = Del_JD[i,j]
    Del[u,2] = Del_highly_correlated_variable[i,j]
    u = u + 1
  }
}
Del=data.frame(Del)
Del = unique(Del)
cor.test(Del$X1,Del$X2)
model.highly_correlated_variable <- lm(Del$X1~Del$X2)
plot(Del$X1~Del$X2,xlab="Cumulative SDE difference (Kg/m2s)", ylab="DoY difference",
cex.lab=1.5, cex.axis=1.25)
grid()
abline(model.highly_correlated_variable,  , col="blue")
summary(model.highly_correlated_variable)
coefficients(model.highly_correlated_variable)

##### Weekend-effect test #####
mydata.individual.year <- which(mydata$Jaar == 2009)
sort(mydata$Jaar[mydata.individual.year])
individual.year <- mydata[mydata.individual.year, ]
individual.year[order(individual.year$nr), ]
nature = ppp(individual.year$x, individual.year$y, window = winn,
unitname=c("metre","metres"),marks=individual.year$DW,check = TRUE)
summary(nature)
fre=data.frame(summary(marks(nature)))
chisq.test(fre, p=c(5/7, 2/7))$p.value

##### Urban-effect test #####
nature = ppp(individual.year$x, individual.year$y, window = winn,
unitname=c("metre","metres"),marks=individual.year$Land_cover,check = TRUE)
summary(nature)
```

```
fre=data.frame(summary(marks(nature)))
chisq.test(fre, p=c(4/5, 1/5))$p.value

##### Complete spatial randomness test #####
nature = ppp(individual.year$x, individual.year$y, window = winn,check = TRUE)
random_VGI = runifpoint(dim[1], win =winn)
P_nature = quadrat.test(nature,method="Chisq", nx=5,ny=6)
plot(P_nature ,cex = 1.25)
plot(nature, add=T)
P_nature$p.value

##### Empty space distance map #####
plot(distmap(nature, eps=1000))
plot(nature, add=T)
plot(winn, add=T)
```